# Use of Databases for Drug Effectiveness Studies

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by Judith Strobl.

June 2010

*Where is the wisdom we have lost in knowledge?*
*Where is the knowledge we have lost in information?*
T.S. Eliot
*Where is the information we have lost in data?*
(as T.S. Eliot might have said)

**Acknowledgements**

My thanks go first to my supervisor, Professor Tom Walley, whose helpful support and steady encouragement has never waned throughout the years. Thanks are due also to my colleagues in the Prescribing Research Group who challenged, supported, and counselled me, particularly Professor Adrian Bagust, and Dr Ian Enzer for his analytical support.

I am heavily indebted to my loving husband, Barry James, whose patience with a wife spending her spare time on a PhD rather than domestic chores must have been tested more than once - although he never showed it!

My wonderful parents Karl und Margarete Strobl deserve my gratitude for having given me most of what really matters in life.

**Declaration**

I am responsible for all the work presented in this thesis, with the following exceptions: Section 4.6 (analysis of Dornase Alfa Case Study), part of section 5.1 (presentation of results of Dornase Alfa Case Study, section 5.1.2 onwards), part of section 5.2 (discussion of results of Dornase Alfa Case Study), and further outcomes data presented as part of Appendix A, which draw on work of a fellow research team member undertaking the detailed analysis of the Dornase Alfa Case Study.

# TABLE OF CONTENTS

# List of Tables

# List of Figures

# Abbreviations

| | |
|---|---|
| ABPI | Allergic Bronchopulmonary Aspergillosis |
| ACE-I | Angiotensin-converting enzyme inhibitor |
| ACR | American Council of Rheumatologists |
| AHRQ | Agency for Healthcare Research and Quality |
| AIDS | Acquired immunodeficiency syndrome |
| ALL | Acute lymphatic leukaemia |
| AMI | Acute myocardial infarction |
| ANCOVA | Analysis of covariance |
| ANOVA | Analysis of variance |
| BARI | Bypass Angioplasty Revascularisation Investigation |
| *B.cepacia* | *Burkholderia cepacia* |
| BMA | British Medical Association |
| BMD | Bone mineral density |
| BMI | Body mass index |
| BPID | Identity number allocated by Quintiles to patients on the ERCF |
| CABG | Coronary artery bypass graft |
| CASS | Coronary Artery Surgery Study |
| CFF | Cystic Fibrosis Foundation (USA) |
| CF | Cystic Fibrosis |
| CG | Control group |
| CGF | Child Growth Foundation (UK) |
| CHF | Congestive heart failure |
| CI | Confidence interval (95%, unless stated otherwise) |
| CRD | York Centre for Reviews and Dissemination |
| CSP | Coordinating System for gaining NHS Permission |
| CXR | Chest X-ray |
| DF508 | delta F 508 (a common gene mutation in CF) |
| DoCDat | Directory of Clinical Databases (NHS Information Centre) |
| DNA | Deoxyribonuclease |
| DQR | Data quality review |
| enrol't | Enrolment |
| ePACT | Electronic prescribing analysis and costs database for the NHS |
| EPOC | Cochrane Effective Practice and Organisation of Care Review Group |
| ERCF | Epidemiologic Registry of Cystic Fibrosis |
| ESCF | Epidemiological Survey of Cystic Fibrosis |
| ESR | Erythrocyte sedimentation rate |
| EU | European Union |
| F | Female |
| $FEF_{25-75}$ | Average forced expiratory flow rate over the middle half of the FVC manoeuvre |
| $FEV_1$ | Forced expiratory volume in 1 second |
| FVC | Forced vital capacity |
| GAO | General Accounting Office (USA) |
| GI | Gastro-intestinal |
| GMC | General Medical Council |
| GORD | Gastro-oesophageal reflux disease |
| GPRD | General Practice Research Database |
| *H.influenzae* | *Haemophilus influenzae* |
| HAQ | Health Assessment Questionnaire |
| HIV | Human immunodeficiency virus |
| HLA | Human Leukocyte Antigens |
| HMO | Health maintenance organisation |
| HRCT | High-resolution computerised tomography |
| Ht | Height |
| HTA | Health technology assessment |
| I.U. | International units |
| IBD | Inflammatory bowel disease |

| | |
|---|---|
| IBMTR | International Bone Marrow Transplant Registry |
| ICD | International Classification of Diseases |
| IQR | Inter-quartile range |
| IRAS | Integrated Research Application System |
| Iv | Intravenous |
| LDL | Low-density lipoprotein |
| LFT | Lung function test |
| LOS | Length of stay |
| LREC | Local Research Ethics Committee |
| M | Male |
| Max | Maximum |
| MESH | Medical subject heading |
| Min | Minimum |
| MRC | Medical Research Council |
| MREC | Multi-Centre Research Ethics Committee |
| N.rep. | Not reported |
| N.s. | Not significant |
| NCI | National Cancer Institute |
| NIGB | National Information Governance Board |
| NIHR | National Institute for Health Research |
| NHS | National Health Service (United Kingdom) |
| NNRTI | Non-nucleoside reverse transcriptase inhibitor |
| NRMI-2 | Second National Registry of Myocardial Infarction |
| NSAID | Non-steroidal anti-inflammatory drug |
| Observ. | Observations |
| Oesoph. | Oespophageal |
| OR | Odds ratio |
| PCP | Pneumocystis carinii prophylaxis |
| PI | Protease inhibitor |
| PIAG | Patient Information Advisory Group |
| Pred. | Predicted (usually in context of % of predicted $FEV_1$ or FVC) |
| PTCA | Percutaneous transluminal coronary angioplasty |
| R&D | Research and Development |
| RA | Rheumatoid arthritis |
| RCT | Randomised controlled trial |
| RR | Relative risk |
| RTI | Respiratory tract infection |
| Rt-PA | Alteplase |
| RV | Residual volume |
| *S.aureus* | *Staphylococcus aureus* |
| SD | Standard deviation |
| SE | Standard error |
| SEER | Surveillance, Epidemiology, and End Result |
| SEM | Standard error of the mean |
| SI | Statutory instrument |
| SPSS | Statistical Package for the Social Sciences |
| Staph. | Staphylococcus |
| TG | Treatment group |
| TLC | Total lung capacity |
| TMP/SMZ | Trimethoprim/sulfamethoxazole |
| t-PA | Alteplase |
| UKCC | United Kingdom Council for Nursing, Midwifery, and Health Visiting (now: Nursing and Midwifery Council) |
| UKCFS | UK Cystic Fibrosis Survey |
| UTI | Urinary tract infection |
| VAS | Visual analogue scale |
| Wt | Weight |
| Yo | Year old |
| %ile | Centile |

# Publications

Strobl J, Cave E, Walley T (2000) "Data protection legislation: interpretation and barriers to research." British Medical Journal **321**: 890-892.

Strobl J, Enzer I, Bagust A, Haycox A (2002) "Quality Assessment of Data from the Epidemiologic Register of Cystic Fibrosis (ERCF) Relating to North of England Centres" *conference poster presentation at the 25th Congress of the European Cystic Fibrosis Society, Genoa (Italy), June 2002.*

Strobl J, Enzer I, Bagust A, Haycox A, Smyth R, Ashby D, Walley T (2003) "Using disease registries for pharmacoepidemiological research: a case study of data from a cystic fibrosis registry." Pharmacoepidemiology and Drug Safety **12**: 459-65.

Strobl J, Walley T (2003) "Published Drug Effectiveness Studies Using Secondary Data Sources" *conference poster presentation at the Annual Scientific Meeting of the Society of Social Medicine, Edinburgh (UK), September 2003.*

Strobl J (2005) "Using Registers - Lessons from a Data Quality Review and Systematic Review", *invited oral conference presentation at the 28th European Cystic Fibrosis Conference, Hersonissos (Crete, Greece), June 2005.*

# Abstract

**Judith Strobl:**
**Use of Databases for Drug Effectiveness Studies**

***Background:***
Health policy decisions need to be based on effectiveness estimates of interventions which are representative of normal clinical practice, in order to determine their likely impact on outcomes and costs in real life. Clinical or claims databases may provide a potential source of relevant unselected data for research. However, effectiveness studies based on such databases experience threats to internal and external validity.

This thesis draws methodological lessons from an attempted effectiveness study of dornase alfa for cystic fibrosis, and expands those with a subsequent comprehensive review of comparable published database studies in order to describe current practice and identify useful approaches. Case study comparisons of observational database effectiveness studies with randomised controlled trials seek to add to our knowledge on the comparability of these two study types and their contribution to an evidence base.

***Aims:***

1. Describe in depth the process of and lessons from conducting a drug effectiveness case study using an existing database (the Dornase Alfa Case Study);
2. Review comparable published studies and explore the features relevant to their internal and external validity;
3. Assess whether drug "Effectiveness Studies based on Databases" (ESDs) show comparable results to those of RCTs when selection criteria are the same and when baseline differences in ESDs have been adjusted for.

***Main findings and conclusions:***
The internal validity of ESDs depends on the valid and reliable measurement and reporting of key variables. The Dornase Alfa Case Study encountered many practical as well as methodological pitfalls. There were significant variations in clinical and recording practices between centres. A detailed data validation exercise highlighted considerable data quality problems relating to key outcome variables (exacerbations and deaths). The sample was too small and the data too imbalanced in terms of enrolment and follow-up to be able to subject them to further multivariate analysis.

The potential strengths of ESDs include greater generalisability, better representativeness of the target population, and a more naturalistic setting. The review of published ESDs showed that these strengths have seldom been demonstrated or exploited in published reports, and hypotheses are not always clearly stated. Also, existing drug ESDs tend not to report issues of data access, management, or quality, and many do not comment on data protection and ethical issues. Significant publication bias is likely. Thus, even ESDs published in high-quality journals fall considerably short of expected quality standards.

Four case study comparisons of drug ESDs and RCTs were hampered because key study characteristics differed between study types. There was no indication that ESDs systematically over-estimated effects compared to RCTs, but different biases left scope for diverging results. Robust implemention of now available guidance on the conduct of ESDs, and of the recommendations arisng from this thesis, as well as further methodological research are vital before ESDs can be more confidently used for effectiveness evidence.

# 1 Introduction

The assessment of the effectiveness of health care interventions relies heavily on evidence generated through randomised controlled trials (RCTs). Over recent years, the values and shortcomings of randomised and non-randomised studies have been debated (Peto *et al.* 1995; Black 1996; Anonymous 2000c; Collins & MacMahon 2001), and methodological research, including several systematic reviews comparing both types of study designs, has been published (Gray-Donald & Kramer 1988; Britton *et al.* 1998; Kunz & Oxman 1998; Benson & Hartz 2000; Concato *et al.* 2000; MacLehose *et al.* 2000; Ioannidis *et al.* 2001).

A random allocation of treatment ensures that treated and untreated patients in sufficient numbers are likely to be very similar in all characteristics other than the treatment intervention. Thus randomisation permits the assumption that any observed differences in measured outcomes are principally due to the intervention, as long as the study is properly blinded and measurements are rigorous (Britton *et al.* 1998). It is generally held that alternative study designs without randomisation (such as cohort, case-control, cross-sectional studies, or non-randomised trials) cannot provide the same degree of assurance in the observed effectiveness results, because any known or unknown differences between patient groups at baseline may influence the observed outcomes.

On the other hand, the validity of randomised studies may also be limited, for example, if blinding of patients and assessors is insufficient or indeed not possible. Randomised studies may be unrepresentative of patient groups not included in such studies. They are also not always possible or practical (Black 1996). The artificial research conditions imposed by a typical randomised trial are unrepresentative of normal clinical practice and care, which has implications for the generalisability of trial results to a wider population. Nevertheless, whereas the advantages of randomisation over no randomisation

are not disputed, the value of non-randomised studies particularly in effectiveness research is hotly debated. Recent reviews showed no clear indication that such studies systematically over-estimate the effects of treatments.

A particular type of non-randomised study, namely those using routinely collected data to evaluate treatments ("Effectiveness Studies using Databases", from hereon abbreviated to "ESDs"), is increasingly used to "*complement or supplement*" (Lewsey *et al.* 2000 p.1) RCTs. This may be used where there is no RCT evidence, because either an RCT would be unsuitable or has not yet been undertaken (to aid the design of a future RCT), or to assess whether the findings from an RCT can be translated into clinical practice. The ready availability of routinely collected data sources, their large and increasing size and comparatively low cost, as well as the availability of increasingly sophisticated analysis methods, make such ESDs attractive as alternatives to RCTs. In addition, ESDs may be considered to have higher external validity than RCTs (GAO 1992). However, the criticisms levelled against non-randomised studies still apply, and data quality and applicability pose additional problems, although these may not be insurmountable. The value of ESDs as a sub-type of non-randomised studies therefore requires further assessment.

The increasing computerisation of healthcare records encourages a wider interest in the use of databases in effectiveness research. This would have the advantage that effectiveness studies could be based on clinical realities, rather than settings which are constrained and distorted by research-conditions. Importantly also, the range and nature of researchable effectiveness questions could expand towards more population health relevant questions. Current evidence - particularly in the pharmaceutical field - is dominated by efficacy questions formulated and investigated by primarily industry-funded RCTs.

There is some guidance on conducting and reporting ESDs (Huston & Naylor 1996; Motheral & Fairman 1997, and more recently von Elm et al. 2007, Gliklich & Dreyer 2007, and Berger et al. 2008); however, there is a lack of in-depth accounts of the experiences of undertaking such studies, as well as of systematic reviews of such studies. This thesis explores the potential value and methodological strengths and weaknesses of drug ESDs, by reporting the in-depth experiences gathered during an attempted ESD, as well as a comprehensive review of ESDs. The importance of this work is underlined by the fact that there is increasing interest in such studies (Rawlins 2008) and the opportunities presented by pre-existing data are rapidly growing - as is the temptation by clinicians and researchers to use pre-existing data for effectiveness studies. Existing reviews include only very few, if any, drug ESDs, and there has so far not been a review dedicated to these studies.

The aims of the thesis are as follows:

1. Describe in depth the process of and lessons from conducting a drug effectiveness case study using an existing database (the Dornase Alfa Case Study);
2. Review comparable published studies and explore the features relevant to their internal and external validity;
3. Assess whether drug "Effectiveness Studies based on Databases" (ESDs) show comparable results to those of RCTs when selection criteria are the same and when baseline differences in ESDs have been adjusted for.

The work began in 1999 with the intention to evaluate the long-term cost-effectiveness of dornase alfa for cystic fibrosis (CF). Dornase alfa is recombinant human deoxyribonuclease (rhDNase), a genetically engineered version of a human enzyme which cleaves extra-cellular deoxyribonuclease (DNA). The drug is administered by inhalation and cleaves DNA in the sputum,

thereby reducing its viscoelasticity. Clearing the lungs of the viscous sputum (usually through physiotherapy) is one of the most important interventions in CF, and dornase alfa is intended to facilitate that process.

At the planning stage for this study, there was evidence for the short-term effectiveness of the drug from randomised controlled trials (RCTs) of up to six months, but the long-term effectiveness, and particularly cost-effectiveness, was under question. Many patients were already using the drug, but clinicians still felt uncertain about the continued benefit of one of the most expensive elements of their treatment regimen.

Given that most eligible patients had already at least tried dornase alfa treatment, future long-term RCTs are prohibitively difficult to initiate in a patient population no longer naïve to the drug. Any potential RCT would have to recruit from a small number of newly diagnosed patients meeting certain eligibility criteria. The loss of equipoise would have been a further reason not to initiate another RCT. It was therefore decided to analyse a subset of data from a large post-marketing surveillance database, the Epidemiologic Registry of CF (ERCF), in an attempt to shed light on the long-term effectiveness of dornase alfa.

In this thesis, I describe the practical experiences of conducting this study in detail. At an early stage, the consideration of then recent UK data protection legislation and guidance posed an unexpected challenge. The process of reaching decisions on access to the secondary data source, and the conflicting guidance and legislation which were available at the time to inform these decisions, are discussed. Based on this experience, I highlighted the problematic implications of the Data Protection Act for medical research in an article in the British Medical Journal (Strobl et al. 2000 - see Appendix A). Secondly, the study incorporated a verification of database records against original patient records in order to assess data quality. This represents the

only direct (and published) verification of data from the ERCF (see Appendix A for a copy of the peer-reviewed publication (Strobl et al. 2003)). Thirdly, an attempt to establish the generalisability of the study is described, and finally the data and initial analyses and analytical challenges are presented and discussed.

The dornase alfa study ran into significant difficulties, not only from a point of view of access to records and data. The detailed considerations of the quality of the available data highlighted more methodological problems in relation to the set up and operation of databases, as well as the use of data derived from them. This resulted in the desire to examine how, if at all, other investigators have managed these problems and what can be learnt from that.

I therefore undertook a comprehensive review of other published drug effectiveness studies based on databases with the intention to examine and describe these studies, their data sources, design, objectives, funding, outcomes, methods, and results, in order to identify design-specific problems and possible solutions. This is the first review focussed on drug effectiveness studies based on databases and its findings are reported here in detail.

Of the ESDs identified in this review, four lent themselves to direct comparisons with RCTs which have addressed the same question. Other investigators have reported systematic comparisons of RCTs and observational study designs, but a systematic approach to such reviews is hampered not least by the comparatively poorer bibliographic indexing of observational studies compared to RCTs. This thesis adds four case studies specifically on ESDs of drug treatments to this body of knowledge. The case studies were selected on the basis of the availability of a Cochrane review having assessed the same question. Study quality was assessed for both ESDs and RCTs, and studies were described in detail before comparisons of results were attempted.

The thesis begins with a detailed review of relevant literature on health technology assessment methods, focusing on RCTs and ESDs (Chapter 2). Much of the empirical work reported in this thesis is exploratory. Both the Dornase Alfa Case Study (Chapters 3-5), as well as the review (Chapters 6) and comparison case studies (Chapter 7) are used to gather insights into the use of ESDs, their intentions, problems, and the value which they may add to a body of effectiveness evidence. Chapters 8 and 9 bring together the findings from these pieces of work and draw out specific lessons for ESD as well as database design, and further areas in need of research. The temptation to use database studies as alternatives to resource-intensive RCTs which ultimately may not be sufficiently generalisable is persistent, as recent debates indicate (Rawlins 2008). This thesis aims to contribute to and further inform this debate.

# 2 Literature Review

## 2.1 Introduction

This chapter provides the context of existing knowledge for the thesis. The relevant fields are charted out, beginning with health technology assessment (HTA) methods, with a particular focus on RCTs and ESDs. It presents a critical review of RCTs as a gold standard for effectiveness evaluations, as well as methodological developments which have moved beyond RCTs, in an attempt to cope with their limitations. In later sections, I review the state of the debate about the value of non-randomised studies for effectiveness research, of which ESDs are an example.

As is customary in multi-topic literature reviews preceding research reports, this chapter organises the material in a "funnel" approach. Hence, it takes the reader from the general to the specific, from health technology assessment in general to the specific open question about the value of ESDs for effectiveness evidence. The identification of relevant literature consequently varied for the different sub-sections of this chapter. Earlier sections are mostly narrative, based on standard texts. The subsequent sections on critiquing RCTs and methodological developments beyond RCTs include increasingly more methodological research and discussion papers identified through a range of Medline searches as well as through "snow-balling" (i.e. texts identified by other authors). The intention is to provide a critical narrative and not necessarily a comprehensive systematic review. Similarly, section 2.5 (Effectiveness studies using databases) aims to be descriptive and illustrative and therefore again uses a limited range of methodological texts. Section 2.6 on previously published systematic comparisons of randomised and non-randomised studies is the most comprehensive and the search strategy is described there.

**CHAPTER OVERVIEW**

## 2.2 Health Technology Assessment (HTA)

This section provides the reader with a brief introduction to Health Technology Assessment and describes the study designs used for determining treatment effect. This section does not lend itself to a systematic review and therefore uses a traditional narrative style. Methodological texts referred to are to a large extent standard textbooks.

The UK HTA Programme defines a health technology as follows:

> *"Health Technology Assessment asks important questions about [health] technologies such as:*
>
> - *When is counselling better than drug treatment for depression?*
> - *What is the best operation for aortic aneurysms?*
> - *Should we screen for human papilloma virus when doing cervical smears?*
> - *Should aspirin be used for the primary prevention of cardiovascular disease?*
>
> *It answers these questions by investigating four main factors:*
>
> - *whether the technology works*
> - *for whom*
> - *at what cost*
> - *how it compares with alternatives"*
>
> (National Coordinating Centre for Health Technology Assessment 2009)

HTA thus includes the assessment of clinical effects, as well as economic evaluations and other non-clinical assessments of, for example, legal, ethical, and social implications of health technologies. For the last group, a large array of study methods may be appropriate, including qualitative interviewing

or observational methods (Mowatt *et al.* 1997). Thus, HTA - like many areas of health-related research - is a field where fundamentally different scientific paradigms meet (chiefly the positivistic paradigm of the natural sciences, and the ethnographic and anthropological tradition of social sciences) and are challenged to integrate conceptually and practically their respective methodological tools. Similarly, the use of epidemiological methods and study designs in HTA is an issue of ongoing debate and development.

### 2.2.1 "Effectiveness"

It is noteworthy that the UK HTA Programme speaks of "effectiveness" rather than "efficacy" (National Institute for Health Research 2009). Thus, HTA is concerned with whether an intervention is effective in practice ("effectiveness"), rather than the mere potential of an intervention to be effective ("efficacy"). The *efficacy* of a treatment concerns its ability to improve or cure certain conditions under optimal conditions (rather than cause harm). By contrast, the *effectiveness* of an intervention concerns its actual effect on patients using or being offered it under real life conditions. Thus effectiveness also considers broader aspects of use in practice, such as the acceptance of the treatment, including compliance (Drummond *et al.* 1997). Related issues of interest to health service planners are *access* to and *availability* of the treatment for those patients who are able to benefit from it.

The United States General Accounting Office (GAO 1992) has described three dimensions of the "effectiveness domain": various types of patients and forms of the disease, varying implementations of the treatment in question, and varying outcome criteria. The effectiveness of any treatment thus depends on all these three dimensions, whereby for many treatments, the size of the effectiveness domain is not known. Thus if a study evaluates a treatment in a

particular respondent group in a particular way, it may be unknown how effective the treatment is in other groups or under different conditions.

Effectiveness information may feed into the evaluation of *efficiency*, or economic evaluation, of health interventions. Such evaluations address the relative value of interventions, i.e. the question of whether they are worth doing compared to other options. The evaluation of different aspects of effectiveness should ideally precede any economic evaluations (Drummond *et al.* 1997). In reality, health economists often draw on a number of studies of different design, using their results to model decision options because of many uncertainties.

## 2.2.2 Study designs for determining effects of treatments

The focus of this thesis is on research designs relevant to the assessment of the clinical effects of health technologies. It is recognized, however, that qualitative and other social research methods can directly inform effectiveness evaluations, for example by exploring aspects of medication use in practice (e.g. dissemination of new technologies, their acceptability to practitioners and patients, patients' motivation to seek or comply with a particular treatment). Before considering the main designs for estimating treatment effects, two important concepts need to be introduced, by which the adequacy of a research design can be assessed: internal and external validity.

### *2.2.2.1 Internal and external validity*

Internal validity of a study is attained if the change in the dependent variable(s) can be said to arise from only the effect of the independent variable of interest (Polit & Hungler 1991). The intention of a study's analysis is to demonstrate that an observed association between independent and dependent variables is not due to either chance, systematic error ("bias"), or

confounding, i.e. an association of any factor with both the independent and dependent variables (Hennekens & Buring 1987).

A range of possible biases may threaten internal validity and can influence study results in both directions (Elwood 1998). Testing and instrumentation effects can be a problem in any before-after study, e.g. if patients are sensitised or there are learning effects. Selection bias is a serious possibility where study subjects in the comparison groups differ in significant factors. In epidemiological studies in particular, recall and observation bias may play a role, where the recall by patients of significant information or the observation methods applied differ between comparison groups (Hennekens & Buring 1987). A further problem may be the misclassification of exposure or outcomes through inaccuracies in data collection.

Confounding - and to some extent selection bias - can be controlled through either study design or analysis by stratification, restriction of the sample, matching, randomisation, or multivariate analysis (Hennekens & Buring 1987; Rothman & Greenland 1998). Of course, this is only possible where the confounding variable(s) are known and measured. Observation bias may be controlled by the blinding of patients and observers, if that is possible. Biases are difficult to control post-hoc, and assessments of biases can often only be undertaken qualitatively and subjectively (Elwood 1998).

External validity refers to the generalisability of a study's findings to other settings or samples. Strictly speaking, this is only possible where a study sample has been randomly drawn from a target population to which findings are then generalisable (provided the study was internally valid). In many cases, studies rely on an accessible patient population, e.g. of patients of particular hospitals or clinicians interested in participating in the study, and authors need to demonstrate the similarity of both the target and accessible populations in terms of their characteristics (Polit & Hungler 1991).

Restrictive exclusion criteria over patient entry into a study, as well as low participation rates can jeopardise external validity (Elwood 1998).

Apart from sample characteristics, the characteristics of the environment or research situation can affect external validity (Polit & Hungler 1991). The Hawthorne effect describes situations where subjects behave differently in a research context from normal life; similarly, subjects may behave differently because the treatment is new, but the effect would not last equally over time. External events may influence the success of treatment, as may the behaviour of the experimenter. Finally, measurement effects may occur which may mean that findings are not replicable in settings where the same set of measurements are not performed. Also, it has to be stressed that external validity without internal validity is meaningless. RCTs should have high internal validity but often suffer from limited external validity.

#### *2.2.2.2 Clinical trials*

Bowling (1997) defines an experiment as follows:

> *"A scientific method used to establish cause and effect relationships between the independent and dependent variables. At its most* basic, *the experiment is a situation in which the independent (experimental) variable is fixed by manipulation by the investigator or by natural occurrence. The* true *experimental method involves the random allocation of participants to experimental and control groups. Ideally, participants are assessed before and after the manipulation of the independent variable in order to measure its effects on the dependent variable". (p.388)*

"A clinical trial is an experiment with patients as subjects." (Rothman & Greenland 1998, p.69). The independent variable in a clinical trial is the clinical intervention whose effectiveness is to be assessed. Subjects in the

experimental group(s) receive the intervention(s) of interest; subjects in the control groups may receive placebo, standard or no specific care.

The key to accomplishing comparability between subjects in different groups with respect to underlying "baseline" characteristics and ensuring that causal inferences can be drawn with a minimum of assumptions, is randomisation. Randomised trials have first been used in agriculture and have been applied to medical research for the first time in the 1940s (Elwood 1998). Random allocation of subjects to the different comparison groups and their resulting comparability with respect to even unmeasured factors ultimately safeguards the validity of the trial.

In addition, many clinical trials are designed to minimise bias through "blinding" (to the assigned treatment) those administering the intervention, patients, as well as those assessing the outcomes of interest. Keeping all these participants in ignorance over the treatment assignment minimised biases introduced through the knowledge of being treated or not. Blinding demands that any real differences between the treatment of subjects in different groups must not be detectable. This can be achieved by identical-looking forms of administration.

Peto and colleagues (1995) argue that randomisation ensures negligible biases and small random errors - both requirements for the detection or refutation of moderate differences in outcomes. By contrast, non-randomised studies frequently have to undertake mathematical adjustments to control for differences between comparison groups. It is often argued that important prognostic factors may be unrecorded or the quality of their recording may vary, and thus an analysis free of bias can never be guaranteed in non-randomised studies (Peto *et al.* 1995).

Peto and colleagues (1995) present four requirements for reliable assessment of moderate treatment effects: *"properly randomised evidence, an "intent-to-treat" analysis of all randomised patients, no unduly data-derived emphasis on particular subgroups, and, finally, an overview of all the relevant randomised trials (without unduly data-derived emphasis on the results from particular studies)"* (p.33). Only large trials are seen as able to avoid moderate random errors. Collins and MacMahon (2001) echo this set of criteria and go as far as claiming that already many premature deaths have been caused by a failure to produce such evidence.

In relation to the fourth stated requirement, much effort has gone into the development of methods for systematic reviews and meta-analyses of RCTs, not least through the international Cochrane Collaboration. Typically, the assessment of the quality of RCTs (and thus their ability to minimise potential biases) plays a central role in systematic reviews, as do assessments of heterogeneity (i.e. impact of potential effect modifiers) and possible publication bias. Non-RCTs are increasingly considered in systematic reviews, but many available reviews and certainly meta-analyses are still exclusively based on RCTs or at least quasi-randomised trials.

#### *2.2.2.3 Observational study designs*

Observational studies do not involve experimentation ("non-experimental studies"); they are, however, also hypothesis-testing studies. Some writers consider both trials as well as observational studies as "epidemiologic studies" (Rothman & Greenland 1998). The two main types of observational study design are cohort studies and case control studies. Either of these could be applied in database studies, so they are briefly described below:

**Cohort studies**

A cohort study is a direct analogue of the experiment (Rothman & Greenland 1998), but the investigator does not assign the exposure/intervention. Rather,

groups are defined (by the investigator) according to their *exposure* to a potential causative agent; hence in a cohort study there is at least one group of subjects who are exposed and one group of unexposed subjects.

**Case-control studies**

A case-control study defines study groups according to the *outcome*, rather than an exposure of interest. Only cases and exposure status are identified, but denominators in each group are not measured. Hence, only *relative* effects can be estimated by case-control studies. These studies are also more prone to bias (Rothman & Greenland 1998).

The lack of randomisation in observational study designs used for the evaluation of intended treatment effects means that there is always a potential for bias due to the non-random allocation of treatment. Several study design features and analytical methods have been proposed and used to control for potential bias (Klungel *et al.* 2004). They include stratification, or matching on certain variables, multivariable statistical techniques, such as linear or logistic regression, or Cox proportional hazards regression (survival analysis), and propensity score adjustment. Klungel and colleagues also review other methods more commonly used in statistics and econometrics. However, there is little empirical evaluation of the validity of such methods.

For database studies in particular, confounding by indication is a considerable problem, whereby patients treated differ in some systematic way from those not treated. This is difficult to overcome in analysis, not least because the basis for the doctor's decision to treat is not always clear and rarely documented systematically in databases.

### *2.2.2.4 Considerations in choice of study design*

Observational study designs are most often used by epidemiologists for estimating the degree of harm caused by an agent. For ethical reasons as well

as issues of study size, trials could not be used to answer such research questions, and trials of beneficial effects are often too small to detect sufficient numbers of adverse events. In health technology assessment, the use of these methods is still controversial. There is a broad consensus, that randomised controlled trials are the "gold standard" at least in that context (Carné & Arnaiz 2000). Other study designs used for efficacy/effectiveness evaluations are allocated "lower" rungs on the quality of evidence hierarchical ladder. Table 2.1 presents an example of such a hierarchy of levels of evidence for therapy/prevention, aetiology/harm, developed at the Oxford Centre for Evidence-Based Medicine (Phillips *et al.* 2001a).

**Table 2.1: Oxford Centre for Evidence-Based Medicine Levels of Evidence for therapy/prevention, aetiology/harm (Phillips *et al.* 2001a)**

| | |
|---|---|
| Level 1a | Systematic review (with homogeneity) of RCTs (i.e. no "worrisome" variations in directions and degrees of results between individual studies; studies with worrisome heterogeneity should be tagged with a "minus" at the end of their designated level) |
| Level 1b | Individual RCT (with narrow confidence interval) |
| Level 1c | "All or none" (i.e. either all patients died before the treatment became available, but now some survive on it, or some patients died before treatment became available, but now none die on it) |
| Level 2a | Systematic review (with homogeneity) of cohort studies |
| Level 2b | Individual cohort study (including low quality RCTs, e.g. <80% follow-up) |
| Level 2c | "Outcomes" research, ecological studies |
| Level 3a | Systematic review (with homogeneity) of case-control studies |
| Level 3b | Individual case-control studies |
| Level 4 | Case series (and poor quality cohort and case control studies)* |
| Level 5 | Expert opinion without explicit critical appraisal, or based on physiology, bench research, or "first principles" |

* Poor quality study: failed to define clearly comparison groups and/or measure exposures and outcomes in the same (preferably blinded), objective way in both groups, and/or failed to identify or appropriately control known confounders (for cohort studies also: and/or failed to carry out a sufficiently long and complete follow-up of patients).

As is implied in the above table, the main shortcoming of observational studies is their proneness to various biases. It has been argued that observational studies therefore have no place in evaluating moderate treatment effects, but are best used for identifying adverse effects or rare outcomes, provided that there are no obvious sources of bias in operation (MacMahon & Collins 2001).

However, from the issues raised above it becomes clear that the choice of study design depends on several issues. For example, a temporal dimension underlies the evaluation of a particular drug; the evaluation questions alter with each phase of drug development and licensing, as well as in the post-licensing period and long-term. Further, the questions different parties need to have answered to inform their decisions vary; for example, licensing authorities require efficacy and safety data relating to the primary effect and licensed use of a drug, whereas physicians may have a range of questions relating to the use of the drug for different conditions, co-morbidities, patient subgroups, durations or dosages. Thus Strom and colleagues (Strom *et al.* 1984; 1985) highlighted that clinicians need more information about the relative efficacy of drugs (relative to other available alternatives) than may be demanded by licensing authorities. The authors demonstrated also that licensed drugs are used for several different indications, often despite a lack of supporting trial evidence.

In a small review Strom and his colleagues (1984) showed that relatively few post-marketing efficacy questions needed to be addressed by experimental techniques. For a sample of 131 potential uses of 100 recently approved drugs, the authors judged (1) whether formal comparative efficacy evaluation was required or whether instead the intended effect would be dramatic enough to be observed from an uncontrolled case series or even a single patient, and (2) whether a non-experimental formal study would involve problems of confounding by indication, and (3) if this was the case, whether the indication could be sufficiently characterised to provide control of confounding through conventional techniques. The judgement was that the absolute efficacy of 89 potential drug uses (68%) could be evaluated from clinical observations alone, and only 6 (5%) required an experimental study design (22% could not be addressed by either design). Even for relative efficacy questions, 94 (72%) were judged to be able to be addressed validly in

a non-experimental study design, i.e. where allocation to treatment occurs for a purpose other than the scientific inquiry. This study is now rather dated, and the authors themselves concede that their theoretical statements that a question would be "studiable" does not prove that a valid study could be performed. More recent experience casts doubts over these assertions on the basis of observational findings having been refuted by subsequent trials (MacMahon and Collins 2001).

Similarly, Black (1996) is frequently quoted in defence of observational studies to evaluate effectiveness of healthcare. He argues that experimental studies may be, (1) *unnecessary* (e.g. where the effect of an intervention is dramatic), (2) *inappropriate* (e.g. for rare events, for long-term future outcomes, or where randomisation influences the effectiveness of the intervention), (3) *impossible* (e.g. when clinicians refuse to participate, ethical objections exist, legal or political obstacles are posed, where contamination is unavoidable, or simply the limits of research capacity are exceeded), or (4) *inadequate* particularly where generalisability of a trial would be poor.

In the evaluation process of new drug treatments, trial designs play an important role, particularly in early phases, up to what is termed "Phase III" studies. These are large-scale RCTs on patients (rather than healthy volunteers) and are usually sponsored by the manufacturers. These trials form the basis of licensing decisions concerning the new drug. Once the drug is licensed, there may be an obligation on the side of the manufacturer to monitor drug safety, particularly given recent high-profile cases of drug withdrawals from the market (e.g. Vioxx). However, companies' incentives to sponsor more effectiveness research may be severely curtailed, unless commercial interests are at stake. It is thus often left to public and research funding bodies to sponsor studies evaluating the new drug in clinical practice.

At the "post-marketing" phase, key questions on long-term effectiveness, or the effectiveness of the new treatment in particular patient groups not included in Phase III trials (e.g. children, older people, and those with co-morbidities) usually remain unanswered. Thus - apart from methodological criticisms mentioned above - these open questions provide further incentives to look for alternative study designs, not least because of the high cost involved in conducting multiple RCTs of the same treatment. Costs may be an even stronger argument, where long-term effectiveness is to be evaluated, or where a substantial proportion of the patient population have already received the new drug and would no longer be available for recruitment into a trial.

## 2.3 Critique of RCTs

I have already indicated that RCTs also present methodological limitations, and this section provides an overview of these.

### 2.3.1 Threats to internal validity

The importance of internal validity to the potential value of a study has already been stressed, and it is a main concern of any study design to safeguard this. As the main sources of alternative explanations of study findings are chance, bias, and confounding, RCTs are in a strong position: randomisation minimises confounding, and blinding minimises bias (Elwood 1998); a typical RCT involves both.

However, Kaptchuk (2001) points out that randomisation eliminates preference, which is a key element of treatment choice. Particularly where a study is un-blinded, "preference bias" may change the observed treatment effect. Cook and Campbell (1979) describe several threats to internal validity, which may occur despite successful randomisation, particularly in cases where blinding is not possible or not maintained. Many of these result from the inevitable inequity due to only some subjects receiving a possibly desirable intervention. Information intended for only one group may "leak" to another group. The authors term this "diffusion or imitation of treatments". Where the comparison group receives no treatment, there is a temptation on behalf of administrators to compensate for their lack of treatment in other ways ("compensatory equalisation of treatments"). Similarly, "compensatory rivalry" by respondents receiving the less desirable treatment is equally possible; these subjects may thus strive to reverse the expected difference. The same respondents may also respond with "resentful demoralisation",

which could impact on the post-test difference in a way that cannot be attributed to the intervention alone.

In drug trials, blinding is relatively more easily achieved and maintained, but there are several reasons why blinding may still fail despite a concealment of the allocation of subjects to different groups and despite the use of placebos. Clinicians or patients may be able to guess the treatment patients are receiving because of the occurrence of particular effects or side effects. Similarly, blinding may be difficult to maintain during treatment withdrawal phases, if experienced clinicians - or indeed patients - are familiar with the typical duration of a trial.

Randomisation as a probabilistic technique relies on chance; thus it is still possible that even substantial differences between groups remain, despite successful randomisation (Elwood 1998). There are analytic techniques to deal with such unequal distributions, but in some cases such unequal distribution may not be known. It is good practice to report the achieved distribution of all factors between the groups.

### 2.3.2 Threats to external validity

Measures which increase internal validity often potentially jeopardise external validity. Thus the RCT with its potentially high internal validity is often criticised for low external validity. Strict criteria may be applied to RCTs to limit the influence of extraneous variation in outcomes and thus ensure that any observed differences between groups are most likely due to the intervention (Simon *et al.* 1994). Such criteria include patient selection criteria, placebo washout phases, or specific treatment and investigational protocols. They result in a homogeneous patient sample with a minimum of co-morbid conditions, maximum compliance, and homogeneous treatment across participating centres, which all reduce variability and may thus

increase the power of the study and decrease the potential for bias (Simon *et al.* 1994). However, the generalisability of such RCTs may be restricted, and examples have been highlighted in many studies (Brown *et al.* 1999; Stegmayr *et al.* 1999; Roy *et al.* 2000; Magee *et al.* 2001).

Others have described cases where an RCT includes a different patient spectrum from that seen in clinical practice as "nonconsent bias" (Kaptchuk 2001). This can arise not only because of restricted sampling, but also because patients participating in an RCT are by default volunteers and may thus be different from the normal patient population. Britton *et al.* (1998) have reviewed the evidence on the external validity of randomised trials and have found that excluded subjects tended to have a worse prognosis than included subjects. Also, less affluent subjects were more likely to be included in treatment trials, but less likely to be participants in a trial of a preventive intervention.

Quite apart from patient selection, intensive treatment protocols pursued in RCTs also jeopardise the external validity of such studies. Similarly, the timeframe of studies may impinge on the ability to project results to other time points. For example, retention rates in short-term studies were found not to be representative of long-term results in clinical practice (Hawley & Wolfe 1991).

### 2.3.3 Threats to validity inherent to features of the RCT

*"Randomly subjecting a person to a milieu of hidden exposures and then spotlighting him or her with relentless observation does not nurture normalcy, nor does it isolate humans from their mental processes."*

(Kaptchuk 2001 p.543)

Kaptchuk (2001) summarises several ways in which the gold-standard design of the RCT can itself give rise to biased results. Firstly, the masking of the treatment may introduce bias by causing feelings of uncertainty which may decrease any real effect; conversely, heightened vigilance may increase the response. Participation in a blinded RCT may cause a range of emotional responses which could influence results. At times these may have differential impacts on treatments and controls (Bergmann *et al.* 1994) and produce what Kaptchuk (2001) describes as a "masking bias".

Similarly, the need to inform patients of possible side effects may significantly increase the likelihood of detecting and reporting them ("consent bias"). Another possible bias is referred to by Kaptchuk (2001) as "investigator self-selection" bias, and he quotes evidence of the possibility of investigators having differential impacts on patients and observed treatment outcomes.

### 2.3.4 Limits to economic evaluations

Health economic evaluations of treatments are based not on the theoretical possibility of the treatment working, but on its likely performance in clinical practice (Rittenhouse 1995; Rittenhouse & O'Brien 1996; Revicki & Frank 1999). These evaluations can be severely misleading, if they are based on RCTs which were limited to a certain patient group or which involved treatment and investigative protocols not normally followed in everyday practice. Critically also, health economic evaluations require data on long-term outcomes, when they examine the discounted costs and benefits, e.g. in estimating cost per life-year saved (Nuijten 1998). Such outcomes are rarely gleaned from resource-intensive and thus time-limited RCTs.

Critics would argue that cost-effectiveness evaluations based on RCT data provide "a very precise answer to the wrong question", and would not represent the effect of a new treatment on costs, healthcare use, and

outcomes in clinical practice (Revicki & Frank 1999). It has been argued that the use of blinding and randomisation distorts the reality which an economic evaluation needs to capture, and that the way treatments are applied in RCTs differs from normal practice (Revicki & Frank 1999).

Thus, even the RCT as the "gold standard" design for effectiveness evaluations shows considerable scope for threats to its validity. The next section explores study designs which may offer options for overcoming some of the limitations of RCTs; in many cases, these design variations allow for preferences of patients and/or clinicians to be taken into account.

## 2.4 Developments beyond RCTs

Several variations of the RCT design have been suggested in the literature in order to overcome some of its criticisms.

### 2.4.1 RCT variants

MacLehose and colleagues offer a helpful overview and discussion of variants of RCTs (MacLehose *et al.* 2000). The first example is the *single randomised consent design* (Zelen 1979), which randomises patients to either a new treatment, or current or no treatment options. Patients are only told about randomisation if they are randomised to the new treatment group. Others are followed up and their data analysed (in an intention-to-treat analysis) without information or consent. Whereas the avoidance of discussing randomisation with patients may sound methodologically attractive, such conduct is now unacceptable in research involving human subjects.

The same author suggests a *double randomised consent design* for situations where there is no control or best standard therapy (Zelen 1981). In this case, patients are randomised to experimental or standard therapy prior to consent. In both groups, patients are then permitted to choose which treatment they wish to receive, but the analysis is again performed by intention-to-treat. Whereas this design is ethically less repugnant, it shares some shortcomings with the *single randomised consent design*; firstly, double-blinding is not possible, and there may be substantial loss of efficiency due to patients choosing the alternative treatment.

*Response adaptive designs* adjust the likelihood of newly recruited patients being allocated to a particular treatment group depending on how successful

the treatment currently appears. Likelihoods are adjusted in favour of treatments appearing more successful, and the exposure to less effective treatments is minimised (MacLehose *et al.* 2000). The complexities in determining likelihoods are, however, substantial, particularly where long follow-ups are required or multiple outcomes are to be assessed.

A further variation of the RCT gives patients the option to change onto open treatment at any point in the study, and analyses the time until this is requested (Hogel *et al.* 1994). One of the advantages seen in the *change-to-open-label design* is that patients might more readily agree to participate, thus making the trial more generalisable. However, for several types of treatment this design seems inappropriate, e.g. where beneficial outcomes are not expected for some time.

Feldman and colleagues (Feldman *et al.* 2001) suggest a randomised study design for situations where it is unacceptable for patients to be allocated to a control group. In the *randomised placebo-phase design,* all patients are randomised to a placebo-phase duration (any number of days), after which treatment begins. The design can thus only assess whether patients treated sooner respond to treatment sooner than those patients commenced later on treatment.

Other variants of RCTs aim to exclude patients who are likely not to comply with treatment through placebo run-in phases (thus increasing the efficiency of the trial), or to minimise the exposure to placebo of patients who are not responding to treatment by eliminating non-responders after a treatment run-in phase (MacLehose *et al.* 2000).

### 2.4.2 Pragmatic or naturalistic trial designs

Over thirty years ago, Lasagna (1974) argued for what he called a "naturalistic" evaluation of medicines. Naturalistic trials aim to utilise the strengths of randomisation and simultaneously to improve generalisability through broad patient selection criteria, long follow-up periods (not least in order to be able to measure clinical endpoints instead of surrogate variables), and substantial flexibility in dosing, prescribing, and care practices. These demands tend to make the trials more expensive, time-consuming and difficult to implement (Carné & Arnaiz 2000) than standard efficacy trials.

Sacristán and colleagues (1998) have described a design which includes randomisation modules into computer-based patient record systems. This method should thus eliminate selection bias and also utilise the strength of large database systems in collecting "naturalistic" data from clinical practice for larger samples and longer study periods. However, the high demands on the degree of computerisation and level of data quality implicit in this proposed design seem to have contributed to its so far limited success. The authors themselves point out possible methodological problems such as cross-over bias, and what Kaptchuk (2001) called investigator self-selection bias.

### 2.4.3 Hybrid study designs

MacLehose and colleagues have as part of their systematic review provided an overview of study designs which provide both RCT and non-randomised effect estimates. An overview of the designs is provided in Table 2.2.

**Table 2.2: Hybrid designs with RCT and non-randomised components**

| Design | Description | Reference |
|---|---|---|
| *Comprehensive cohort study* | Patients who refuse randomisation may choose any comparison treatment and are also followed up so that generalisability of the RCT can be tested | (Olschewski & Scheurlen 1985) |
| *Patient preference trial* | Patients with strong preferences for one of the comparison treatments (for whatever reasons) are allocated the preferred treatment, indifferent patients are randomised; effects in randomised and non-randomised groups are compared. | (Brewin & Bradley 1989) |
| *Clinician-preferred-treatment trial* | Patients for whom either comparison treatment is clearly indicated receive the appropriate treatment; where the indication is less clear, a panel of clinicians can either agree on a preferred drug for each individual patient, or if the panel does not agree, patients are offered randomisation to either treatment and then treated by a clinician with a preference for the randomly allocated treatment; effects in randomised and non-randomised cohorts are compared. | (Korn & Baumrind 1991) |
| *Two-stage trial design* | Patients are randomised to a patient-preference trial or to a further randomisation into an RCT | (Rucker 1989) |

Source: MacLehose *et al.* (2000).

## 2.4.4 Research syntheses

The need to summarise the available evidence for decision making in healthcare and medicine has given rise to new quantitative synthesis techniques (Petitti 1994; Labin 2007). The evidence-based practice movement and HTA programmes in several countries have contributed much to the development of research synthesis techniques in HTA. In particular, the review work of the international Cochrane Collaboration centres on meta-analytic techniques. National and international programmes and collaborations exist to coordinate and undertake systematic and comprehensive syntheses.

Research syntheses depend firstly on a comprehensive search and identification of all relevant available studies. Without such systematic

literature reviews, bias is almost inevitable. Search criteria and data extraction and handling procedures are to be defined a priori, based on a clinical question, rather than previous knowledge of the available literature. Manipulating inclusion criteria can bias a review (Jüni *et al.* 2001). The identification of RCTs has recently been made easier through concerted efforts in registering trials and coding trial publications (Thompson 2001), but search strategies for identifying non-randomised studies are much more difficult to design (MacLehose *et al.* 2000).

Despite systematic searches, potential biases remain. Firstly, publication bias poses considerable problems, not only where industry-sponsored trials are predominant, but also where researchers fail to submit or editors fail to select for publication those trials which do not show statistically significant results. This means that published studies may differ systematically from unpublished studies, and both should therefore be included in a systematic review (Thompson 2001). Similarly, the identification of studies published in less widely-used languages is difficult, but more importantly, such studies may differ systematically from studies published in English (Jüni *et al.* 2001). Trial authors may jeopardise the validity of a subsequent review through their choice of outcomes reported, other studies they cite, and even number of publications they produce based on the same study.

The quality of evidence from any research synthesis is only as good as the quality of the studies and information on which it is based. The quality of evidence is difficult to define, and may address the design, implementation, and/or reporting of a study (Jüni *et al.* 2001). Again, the methodological developments for assessing the quality of studies are most advanced in the case of RCTs, for which validated quality assessment instruments exist (Jadad *et al.* 1996). Whereas a number of tools for assessing the quality of observational studies are available, many are not fully developed, and a consensus on the critical elements for assessing the susceptibility to bias is

still outstanding (Sanderson et al. 2007). The CONSORT group guidelines for the reporting of observational studies (von Elm et al. 2007) could be a starting point.

#### *2.4.4.1 Meta-analysis*

A meta-analysis is a statistical method for combining results from a number of studies. Apparently the term was coined by Glass in 1976, but appropriate statistical methods had already been developed in the 1930 (work by Tippett, Fisher, Cochrane, and Pearson), before being widely used in the social sciences in the 1970s (Petitti 1994). Whereas most meta-analyses are limited to RCTs, there are increasingly attempts to include non-experimental data in meta-analyses.

The result of a meta-analysis is an overall estimate of effect of an intervention, based on a number of trials. However, different trials included may be heterogeneous, and an overall result of the meta-analysis may not apply to particular subgroups of patients or treatment practices. An exploration of any potential sources of heterogeneity in a meta-analysis is therefore important and can be very informative. Meta-regressions can assess how trial or patient characteristics can influence treatment effects (Petitti 1994).

The limitations of meta-analyses in terms of their contribution to healthcare decision-making relate to the reliance on predominantly RCTs. Meta-analyses thus suffer from the shortcomings of RCTs, mainly in terms of their lack of generalisability. Poor reporting of trials poses an additional limiting factor, as many older trials do not provide the necessary detail (point estimate plus standard error or confidence limits) to be included in a meta-analysis (Greenland 1987). As researchers retire, this information will be increasingly difficult to retrieve retrospectively.

Evidence-based decision-making does not exclusively rely on meta-analyses, as (clinical) experience, contextual knowledge, and other types of information are taken into account. Thus, there is probably more readiness on the part of decision-makers than statisticians to ignore the findings of a meta-analysis whose main outcome is a recommendation for further research.

It is possible in meta-analyses to examine subgroups of patients for differential responses. However, it has been suggested that such probing effectiveness analyses can be misleading and therefore are to be seen as exploratory (Temple 1999). Such analyses need to account for multiple comparisons by specifying a-priori hypotheses and correcting p-values accordingly in order to avoid spuriously significant findings.

Berlin and Colditz (1999) suggest that meta-analyses could be pre-planned by specifically developing a series of trials on different patient groups; the final meta-analysis of such trials should thus provide better generalisability as well as the possibility to identify variations between subgroups of patients. Temple (1999) argues that, multi-centre trials might serve this purpose. Further, he suggests that if such meta-analyses were to examine defined differences between settings, each individual trial would have to be large enough to show an independent effect. A meta-analysis would be valuable for such a case only if individual studies are not large enough to examine further effects or outcomes of interest (Temple 1999).

#### *2.4.4.2 Broader-based synthesis methods*

There is a growing methodological literature on statistical methods to include evidence from other study designs (Eddy *et al.* 1992; Sutton *et al.* 1998). In the 1980 and 1990s, the United States Government Accountability Office (formerly General Accounting Office) has developed several synthesis methods with a view to integrating different research methods to answer broader-based policy questions (see Labin 2007).

With particular relevance to ESDs, the then General Accounting Office (GAO 1992) proposed "cross-design synthesis" as a means for improving knowledge about the effectiveness of interventions by synthesising findings from complementary research designs. Labin (2007) describes this as a "bridge between meta-analysis and the evaluation and prospective synthesis methodologies from the GAO." Results from studies with designs of complementary strengths and weaknesses are to be combined, after a systematic and detailed assessment and taking account of different studies' weaknesses in the synthesis of the results.

The GAO (1992) describes the steps for the example of a synthesis of results from randomised studies and database analyses, which are considered complementary designs insofar as they have opposite strengths and weaknesses in terms of their internal and external validity. Whereas RCTs have the potential to ensure high internal validity, they often sacrifice external validity. The opposite may be true of ESDs which may be criticised for their potential selection and other biases, but can have significantly higher generalisability due to their inclusiveness and large patient numbers. The four major steps of cross-design synthesis of these study designs have been described as:

- assessing randomised studies for their generalisability to all relevant patients
- assessing ESDs for "imbalanced comparison groups"
- adjusting individual studies' results compensating for biases if necessary
- synthesising adjusted results within and across design categories.

(GAO 1992)

This approach has since been built on in further developments of broad-based syntheses (Greenhouse & Kelleher 2005; Sutton et al. 2000).

### *2.4.4.3 Decision analysis*

Decision analysis uses quantitative methods to compare the values of different decision options. The method originates in game theory, and in the medical context is used to assist either decisions about the management of individual patients, or - increasingly - the development of health(care) policies (Petitti 1994). The identified "problem" has to be structured and broken down into components and decision options, represented in a decision tree. Outcomes are defined, and uncertainties identified. Values of outcomes are measured or inferred, and uncertainties estimated on the basis of literature reviews and expert opinions, before the decision tree is analysed statistically to estimate the relative net value of the decision options (Pauker & Kassirer 1987).

While intuitively appealing, the methods can be complex and based on many assumptions. For example, Markov modelling, which is used to represent transitions in and out of various states of health, assumes independence of subsequent transitions from previous ones. Also, data to estimate transition probabilities may not be available. Clear descriptions of methods and assumptions as well as sensitivity analyses are thus vital.

### *2.4.4.4 Cost-effectiveness analysis and other economic evaluations*

Ultimately, HTA informs not only decisions at individual patient level but also at the level of entire patient populations and finally whole populations for whom healthcare options are chosen. Economic evaluations are thus an increasingly important part of HTA. It has already been mentioned that health economists have voiced reservations about basing their evaluations on clinical trials alone (Rittenhouse & O'Brien 1996). In order to provide decision-makers with the necessary evidence on the economic consequences of possible decision options, health economists build on effectiveness evidence from an increasing variety of study designs.

In summary, alternative methods to RCTs focus in part on a combination of strengths from various study designs, as well as methods for synthesising evidence to answer questions to support individual or population treatment decisions in practice. Experience with some of these study designs however is still limited. I will now focus on ESDs and their use in effectiveness research.

## 2.5 Effectiveness studies using databases

### 2.5.1 Introduction

Secondary data sources have increasingly been used for health services research, including effectiveness research (Antczak-Bouckoms *et al.* 1991a), despite continued controversy about the quality and validity of evidence so derived (Byar 1980; Shapiro 1989; Antczak-Bouckoms *et al.* 1991a). This section reviews more closely the different types of secondary data sources used for effectiveness research, as well as the challenges and problems associated with such analyses.

### 2.5.2 Secondary data sources

Huston and Naylor (1996) identify two types of secondary data. Firstly, there are databases designed for ongoing surveillance of care. Examples of such databases are post-marketing surveillance databases, but also clinical databases or disease registries, which are focussed on a particular patient population, often those suffering from a chronic condition. Secondly, databases may be designed for administrative purposes, such as hospital databases, or claims databases of insurance companies or health care maintenance organisations. Similarly, administrative databases may have as their subject matter the long-term care of a patient population (Berlowitz *et al.* 1997). Both types of data sources are described further below. Any of these data sources may be used for the evaluation of health interventions. In addition, vital statistics and demographic data (Blais 1991) and also databases generated from large RCTs (Canto *et al.* 1999) have been identified as possible data sources used in health technology assessment and outcomes research.

### *2.5.2.1 Surveillance databases*

Surveillance has been defined as "a continuous and systematic process of collection, analysis, interpretation, and dissemination of descriptive information for monitoring health problems" (Buehler 1998). This definition implies that the main objective of surveillance is health monitoring. There is no consistent use of terminology regarding different types of surveillance data sources. What some describe as a disease registry, others may term a database, a post-marketing surveillance programme or system, or loosely a cohort study.

Buehler defines registries as "listings of all occurrences of a disease, or category of disease (e.g., cancer, birth defects), within a defined area" (Buehler 1998, p.450). Prime examples of such registries are the regional and national cancer registries where all occurrences of malignant neoplasms are registered in the UK, or the Surveillance, Epidemiology, and End Result (SEER) project of the National Cancer Institute (NCI) in the USA (Klawansky *et al.* 1991b; Wyshak *et al.* 1991).

Disease-specific registries may vary depending on their prime intention. They often go far beyond listing disease occurrences. Some registries are designed to follow a patient group over time, usually with the aim to determine the incidence and natural course of a disease, their survival, or incidence of adverse events, as in the case of post-marketing surveillance. Patients are registered on first entry to the database and follow-up information can be entered either periodically (e.g. on regular review visits of patients) or on the occurrence of specific events. Patients are usually identified through clinics or hospitals on the basis of their specified diagnosis, e.g. rheumatoid arthritis (Singh 2001), cystic fibrosis (Anonymous 1994), acute myocardial infarction (Every *et al.* 1999), human immunodeficiency virus (Ledergerber *et al.* 1994; Tassie *et al.* 1999) or a particular intervention, for which follow-up data is

collected (e.g. International Bone Marrow Transplant Registry (Rimm *et al.* 1991).

Disease-specific surveillance systems include the ARAMIS (Arthritis, Rheumatism and Aging Medical Information System) programme, which is intended as a post-marketing surveillance system to explore effectiveness, as well as safety of therapies for rheumatic diseases (Singh 2001). ARAMIS is sponsored by the US National Institutes of Health, with supplemental funding from a range of pharmaceutical companies and the Food and Drug Administration.

A further example of a post-marketing surveillance database is the Epidemiologic Registry of Cystic Fibrosis (ERCF), which provided the data for the Dornase Alfa Case Study presented as part of this thesis (see section 4.1 for more details). The stated aims of this registry include the assessment of effectiveness of a particular treatment (dornase alfa) marketed by the pharmaceutical company funding the registry (Anonymous 1994). A sister database to this exists in North America, the Epidemiological Survey of CF (ESCF).

Some post-marketing surveillance involves very short follow-up. An example of a cross-sectional registry practically without follow-up is the National Registry of Myocardial Infarction (NRMI-2) in the US. It has multiple aims, including descriptive epidemiology as well as safety monitoring of alteplase, a thrombolytic agent (Every *et al.* 1999). This registry collects cross-sectional data on patients admitted with myocardial infarction in the USA and is also funded by a pharmaceutical company (Genentech). A similar database is the Cooperative Cardiovascular Project by the Health Care Financing Agency, albeit that patients discharged from hospital after an acute myocardial infarction are identified centrally for this database (Every *et al.* 1999).

In some cases, healthy volunteers may be recruited to longitudinal cohort studies. The data from these cohorts of individuals may be used to examine several hypotheses relating to normal changes occurring with age, or the development of risk factors. For example, the Framingham Heart Study selected a random population sample of adults (to which further volunteers were added later) beginning in 1949, to examine the development of risk factors for coronary heart disease (Roberts *et al.* 1991). The Veterans Affairs Normative Aging Study was initiated in 1963 to observe health changes through the aging process in healthy men, who are followed up with examinations and tests every five years (Antczak-Bouckoms *et al.* 1991b).

Researchers at the London School of Hygiene and Tropical Medicine have systematically gathered detailed background information on clinical databases existing in the UK. This Directory of Clinical Databases ("DoCDat") is available to be reviewed online (now through the NHS Information Centre). The intention is to provide details on coverage, variables, as well as data accuracy to potential data users. In 2002, 44 clinical databases were registered (the ERCF was not among them); in March 2009 the DoCDat website listed 158 databases.

#### *2.5.2.2 Administrative databases*

Databases compiled for administrative - often billing - purposes have also been used for research, including effectiveness and outcomes research; they can of course also be used for surveillance purposes (Buehler 1998). They are, however, not designed for this purpose. Typically, such databases hold information relating to resource use of individual patients, e.g. investigations, treatments, hospital admissions and discharges, diagnoses, operations, and reimbursed prescriptions.

Information technology has enabled the increasing computerisation of health records, and hospital and health service administration systems. In the UK,

VAMP and AAH-Meditel were the two main commercial organisations responsible for computerisation of general practices (Currie & MacDonald 2000). They have fed into two large datasets, which were widely used for post-marketing research: the UK General Practice Research Database, and MediPlus, a commercial database of Intercontinental Medical Statistics (IMS, UK and Ireland). Both also include clinical data.

Compared with many surveillance databases, administrative data sources may contain less detailed clinical data, making effectiveness evaluations more difficult. Thus the opportunities for record linkage between data sources is particularly important, given that datasets tend to be subject-specific and by themselves rarely contain sufficient information to examine effectiveness-related hypotheses (Klawansky *et al.* 1991b). In the UK, a small number of record-linkage systems exist, e.g. the Oxford Record Linkage Study, the Scottish Medical Record Linkage System, and the database of the Medicines Monitoring Unit at Dundee University (Currie & MacDonald 2000). In other countries, more extensive systems are available (Currie & MacDonald 2000).

Some Health Maintenance Organisations, as well as the USA Department of Veterans Affairs hold particularly comprehensive health information systems relating to subscribed individuals. One of the most impressive initiatives, however, is the healthcare information system for the province of Saskatchewan in Canada, which covers the entire population of 1 million inhabitants. Several separate databases on e.g. hospital discharges, prescriptions, birth and death certificates etc. can be linked for drug surveillance, health outcomes research and other health services research. These databases are regularly used for epidemiological research (Saskatchewan Health 2005).

### 2.5.3 The use of databases for effectiveness evaluations

Alpert (2000) lists several possible uses of clinical registries, including the provision of comparative pictures of the progress of a disease and its management, or information on the outcomes or resource utilisation associated with the disease. Effectiveness analyses do not feature on this list. However, secondary data sources are being used for health technology assessment, despite this not being their (main) objective (Blais 1991). Effectiveness evaluations thus constitute only a minor, albeit growing, proportion of the analytical output of clinical or administrative databases (Wyshak *et al.* 1991).

Work on the use of secondary data in effectiveness analysis published in a special section of the International Journal of Technology Assessment in Health Care in 1991 ("The contribution of medical registries to technology assessment") highlighted several ways in which secondary data can contribute to technology assessment (Klawansky *et al.* 1991a; Roberts *et al.* 1991):

- Evaluate technologies reported there and received by registered individuals to a varying degree;
- Evaluate technologies not reported there or used in clinical practice, but administered to registered individuals (or their data or stored serum) for research or evaluation purposes;
- Evaluate technologies unrelated to the registered population sample (e.g. applying risk predicted by logistic regression equations from the database to other samples);
- Use of registered individuals as control group in comparison with various experimental groups;
- Assessing selection bias in clinical trials;
- Facilitating evaluations of classification and coding systems.

Sørensen and colleagues (1996) provide a useful summary of factors affecting the value of secondary data in epidemiological research (Table 2.3). Whereas this provides a guide for ESD researchers, it can also support the evaluation of published ESD reports.

**Table 2.3: Factors affecting the value of secondary data in epidemiological research (Sørensen *et al.* 1996, p.436):**

1. Completeness of registration of individuals
    a. comparing the data source with one or more independent reference sources
    b. comprehensive records review
    c. aggregated methods
2. Accuracy and degree of completeness of variables
    a. Precision
    b. Validity
3. Size of the data sources
4. Registration period
5. Data accessibility, availability and cost
6. Data format
7. Record linkage

Several recent reviews funded by the UK HTA Programme have explored the use of routinely available data in HTA, with only few of them addressing effectiveness research. One of the first (Lewsey *et al.* 2000) was dedicated to exploring the use of routine data to complement or supplement RCTs. Three case studies of surgical interventions were presented to illustrate the use of routine data in situations where (1) an RCT was not feasible, (2) two established surgical interventions were compared, and (3) trial results were compared with those achieved in routine practice. Although the first case study did not attempt to provide an alternative to an RCT, its findings are relevant: namely that neither scope nor quality of routine data were sufficient to allow such studies, but that changes to data collection could make them possible at least in theory. For example, the lack of data on complexity and severity of a condition of interest should be remediable. This finding was

echoed in the two other case studies, which more directly investigated effectiveness questions, albeit that the scope and quality of the data varied depending on the clinical topics studied.

A review by Williams *et al.* (2003) set out to estimate the feasibility, utility and resource implications of electronically captured routine data for HTA by RCTs. This was done by replicating the analysis of four RCTs by simulating them using routine data. For each RCT, analyses based on designed data were compared with an analysis based on routine data. The research team concluded that routine data could answer the questions posed in RCTs, as well as clinical effectiveness questions by using available proxy measures. Despite this encouraging finding, the authors point out that the validity of routine data and the practicalities of using them still pose considerable problems in reality. The authors recommended prospective testing of the use of routine data in HTA RCTs.

Raftery et al. (2005) examined the use of 270 known English databases more generally and assessed their potential use in HTA (examined were those existing by the year 2000). Twelve of them were reported to have been used for effectiveness research, and a further 10 for comparative audit. However, only very few of the referenced studies were drug effectiveness studies. The authors reported that key aspects of the validity of these databases were often difficult to assess (e.g. completeness of variables or registration of individuals, and that validations of data accuracy was very uncommon), and concluded that few existing databases met the required criteria for effectiveness studies.

In May 2007, the American Agency for Healthcare Research and Quality published a guide to the use of registry data in evaluating patient outcomes. The guide describes what is considered good practice in creating, operating, and evaluating registries, and using them in research (including effectiveness

research); it does not review the quality of such studies (Gliklich & Dreyer 2007).

Increasing numbers of database or registry analyses are being published from a variety of clinical fields and data sources, and the debate about the use of registries or databases for the evaluation of treatment effectiveness is still very much alive (Rawlins 2008), albeit that there are few recent methodological publications evaluating or critiquing such research in practice.

### 2.5.4 Advantages and disadvantages of effectiveness analyses using databases

Firstly, a greater generalisability of ESDs is often held up as one of the advantages over randomised controlled trials, as large databases can cover a more diverse patient sample (GAO 1992). RCTs tend to apply strict and narrowly defined inclusion and exclusion criteria, which limits the ability to generalise findings. For example, Pincus (1988) notes that fewer than 25% of consecutive patients with rheumatoid arthritis are eligible for participation in most clinical trials.

However, claims databases in particular may be limited to a particular (insured, treated) patient group, which may not be representative of the overall patient population (Motheral & Fairman 1997). Similarly, regional practice patterns may vary. Lewis (2001) has criticised cystic fibrosis (CF) registries for a lack of representativeness of the target population, as incomplete and poorly defined sub-populations are covered in such registries. Moreover, there is a strong cohort effect in cystic fibrosis, as previously few cases survived to adulthood. Thus, adult "survivors" currently registered will be different from future adults with CF.

However, the desirability of a more comprehensive population coverage and thus better generalisability of many databases compared to available randomised studies is for many unquestioned (GAO 1992). The representativeness of a database analysis of course depends on inclusion and exclusion criteria applied to data on individuals in the database, as well as those applied for the ESD itself. If too many cases are excluded, the results of an ESD may be no more representative than those of a randomised controlled trial (or may just represent a different sub-population).

A second potential advantage of ESDs is the long duration of patient follow-up in administrative and disease databases. This is comparatively difficult and costly to achieve in randomised controlled studies. Researchers investigating a particular hypothesis do not have to plan and execute complex data collection operations, but can use readily available data and thus arrive at results - even for longitudinal evaluations - much faster which of course makes an individual study much less costly (Lewis *et al.* 1993). A further advantage of ESDs is the possibility to link data from different datasets for investigating a range of different hypotheses not necessarily anticipated at the initiation of the data sources.

Researchers of interventions for chronic diseases have been particularly keen on disease database analyses. The lack of follow-up and real-life context is often criticised in RCTs. Starmer and Lee (1982) note that the emphasis of RCTs is on patient groups rather than individuals, whereas clinicians are drawn towards more detailed subgroup analyses in order to inform individual patient decisions. Crucially, they note that "*once one embarks on multiple comparisons outside the design hypothesis, accurate estimation of inferential error rates is lost in a randomised design*" (Starmer & Lee 1982, P.1079). RCTs are usually too small and too brief to handle this problem, or indeed the issue of periodic changes relevant to HTAs in chronic disease areas.

One of the strongest arguments in favour of ESDs is the claim to a real-life naturalistic environment in which patients receive their treatment, regardless of and even oblivious to any treatment evaluation undertaken on the collected data (Lewis *et al.* 1993). Thus the use of routine data sources can eliminate or at least reduce certain biases such as patient selection bias, recall bias, or volunteer bias (Sackett 1979).

However, substantial room for bias and confounding remains in ESDs. The main dangers to the validity of an effectiveness study are very similar to those faced in non-randomised studies, particularly confounding by indication, which may leave even experienced analysts unable to draw sensible conclusions from the data (Byar 1980). Motheral and Fairman (1997) provide a comprehensive overview of possible limitations of the use of claims databases in outcomes research. As threats to internal validity, they mention problems with coding diagnostic information, poor reporting of compliance, misclassification of exposure, referral bias, protopathic bias (i.e. change in treatment because of the baseline manifestation caused by a disease or other event), and confounding by risk factor, severity, or indication.

In order to control for confounding by indication, the data source needs to contain sufficient detail to allow an assessment of the possible confounding variables at baseline in those exposed and those not exposed to the intervention of interest (Roberts *et al.* 1991). Propensity score analysis methods have been suggested to assess a database's ability to address a particular hypothesis of causal effects (Rubin 1997), and other statistical modelling techniques can be used to adjust for confounders (e.g. Cox proportional hazards model). However, it can be argued that it is impossible to rule out residual confounding by unobserved variables (Green & Byar 1984).

Other more practical - and theoretically at least partly avoidable - problems include inconsistent reporting or coding of information in databases, changes

of definitions over time, and a lack of information about those changes, missing data, and defining "time zero", particularly for patients with chronic diseases (Byar 1980; Shapiro 1989; Byar 1991). Even more problems arise in prescription databases which often contain limited if any diagnostic information, e.g. of chronic and secondary conditions or the timing of a diagnosis (Lewis *et al.* 1993). Drug prescriptions alone are not a sufficiently rigorous definition of exposure (Shapiro 1989) and may pose a threat to construct validity, e.g. where a drug can be used for more than one indication (Motheral & Fairman 1997).

### *2.5.4.1 Recommendations for ESDs*

Against the background of continued controversy about the validity and usefulness of database analyses, some authors have published helpful suggestions and recommendations for conducting and interpreting ESDs (Huston & Naylor 1996; Motheral & Fairman 1997; von Elm et al. 2007; Gliklich & Dreyer 2007):

The paper by Huston & Naylor (1996) in the Canadian Medical Association Journal provided an early helpful set of guidance on reporting studies based on secondary data analyses. Study authors are advised to pay particularly close attention to the assessment of any possible random and systematic errors. In addition, the level of detail in the description of the manipulation of the database should be such as to allow replication of the reported study. It is suggested that authors need to defend their choice of data source by clearly demonstrating that it does indeed capture the variables of interest to the study.

This advice is echoed by Motheral and Fairman (1997) in their paper on the use of claims databases in outcomes research. Their advice is more methodological and comments on the need for a comparative study design, and for ensuring that study design and objectives need to be compatible with

the database. The authors also advise researchers to assess carefully, which variables their study needs to have available, and whether these are valid, reliable, and complete in the database; relationships between variables need to make clinical sense. For example, if a confounding factor is known, its assessment and adjustment must be built into the study. Lastly, the authors recommend sensitivity analyses to assess the impact of methodological decision, and the reporting of all relevant information important for the interpretation of the study and its generalisability to a particular context.

More recently, von Elm et al. (2007) have published a consensus guidelines on the reporting of observational studies more generally, which is eminently applicable to ESDs. In the same year, the US Agency for Healthcare Research and Quality has published its guidance on "Registries for Evaluating Patient Outcomes". In over 200 pages, this publication covers everything from planning and design of registries to data collection and quality assurance (Gliklich & Dreyer 2007).

Thus, plenty of advice exists now for ESD researchers, but I am not aware of any reviews of whether published ESDs heed this advice and how they overcome potential methodological pitfalls.

## 2.6 Published systematic comparisons of randomised and non-randomised studies

This section reviews the current literature on systematic comparisons between randomised and non-randomised studies of the same intervention. Comparisons of RCTs and non-randomised effectiveness studies of single interventions were excluded. Reviews were identified from a Medline search including relevant methodological terms describing both RCTs and observational studies (previously used by Britton *et al.* 1998). The search was undertaken in 2001, a time when the key studies by Benson and Hartz (2000) and Ioannidis *et al.* (2001) had just been published. The references of these studies and several others discussing them at the time were examined. In addition, the electronic table of contents of key journals has been scanned for further studies until August 2007 (International Journal of Epidemiology, The Lancet, Journal of the American Medical Association, New England Journal of Medicine, Journal of Clinical Epidemiology, Journal of Epidemiology and Community Health, and British Medical Journal). Given the lack of appropriate indexing of relevant reports in bibliographic databases, this review cannot be entirely systematic, but it draws on the key papers available at the time and since.

The underlying assumption in such comparisons is that observational studies are to be validated against the randomised controlled trial. This may not be entirely justified. Kaptchuk (2001) argues that the appeal of the randomised controlled trial was never based on empirical evidence. He gives a historical overview of the development of the widely held view that non-randomised methods produce higher estimates of outcomes and favour new treatments. However, recent systematic reviews have seriously undermined this assumption.

Kunz and Oxman (1998) have examined the relationship between randomisation and effect estimates in a systematic review of four types of comparisons: RCTs versus non-RCTs of the same interventions, RCTs versus non-RCTs across different interventions, adequately concealed versus inadequately concealed random allocation in trials, and high quality trials versus low quality trials in which the effect of randomisation could not be separated from the effects of other methodological manoeuvres. No comparisons between RCTs and observational study designs were made.

In the first comparison, eight identified comparisons represented both under- and over-estimations of effects in non-RCTs compared to RCTs. In the second type of comparisons, which included three reviews, substantial heterogeneity and other potentially distorting factors made the interpretation difficult. No consistent relationship between study design and effect size were found. The third comparison group only included two studies, both of which showed an over-estimation of effect in studies with inadequately concealed allocation. The fourth comparison (high versus low quality trials) found considerable differences in effect estimates with distortion in both directions. The authors conclude that "*the unpredictability of randomisation is the best protection against the unpredictability of the extent and direction of bias in clinical trials that are not properly randomised*" (Kunz & Oxman 1998, p.1189).

Two reviews undertaken through the UK HTA programme have examined the issue (Britton *et al.* 1998; MacLehose *et al.* 2000). Reeves and colleagues (1998) hold that "*the perception that non-randomised estimates consistently favour new treatments has led to the assumption that discrepancies arise from bias*" (p.73). Thus the first review (Britton *et al.* 1998) focussed on issues of external validity, whereas the second (MacLehose *et al.* 2000) addressed issues of internal validity.

Britton *et al.* (1998) identified 18 papers reporting evaluations of the same intervention by non-randomised studies and RCTs. Overall, there was no consistent discrepancy between the effect estimates of studies of both designs. The authors identify several reasons for discrepancies in either direction, which could be due to chance, differences in study populations, timing, or the nature of the intervention. The review found that the results of both designs concur best if the same exclusion criteria are applied and potential confounders controlled for in non-randomised studies.

The second HTA programme review (MacLehose *et al.* 2000) compared randomised and non-randomised studies of the same intervention reported in the same paper. In a second comparison, two interventions were compared, regardless of the source of their reports. These interventions were mammography screening to reduce breast cancer mortality, and folic acid supplementation for women trying to conceive to prevent neural tube defects in babies. These interventions were chosen on the basis that the nature of the intervention, the populations, and outcomes studied would be similar across studies.

For the fourteen papers reporting direct comparisons (38 comparisons), MacLehose *et al.* (2000) judged the fairness of these comparisons based on the comparability of the studies and the susceptibility of the comparisons to bias. For comparisons judged as "fair", discrepancies between the results of both designs were found to be not significant and were much smaller than for "unfair" comparisons. For the latter there was a tendency in non-randomised studies to show more extreme results than randomised studies, particularly for large sample sizes.

In the comparisons of studies on mammography screening and folic acid supplementation, effect size estimates of RCTs and cohort studies were similar, but case-control studies gave significantly different effects estimates.

However, the direction of the discrepancy could not be predicted, and the authors suggest that it could be specific to the type of intervention.

In June 2000, the New England Journal of Medicine carried two further reviews of randomised versus observational studies. Concato and colleagues (2000) identified RCTs and observational studies from published meta-analyses and were able to perform comparisons of the original studies on five topics (99 reports). For each topic and study design, summary estimates and confidence intervals were determined. This review found that well-designed observational studies showed very similar results to RCTs and also showed less variability in point estimates than RCTs. The authors challenge the often awarded higher status of RCTs in the hierarchy of research designs.

Benson and Hartz (2000) searched first for observational studies comparing two or more interventions for the same condition. Subsequently, trials and further observational studies of the same research question were searched for. Results were pooled by research design (randomised or observational) for each of 19 treatment comparisons (total of 53 observational studies and 83 RCTs). The combined effect size of observational studies lay outside the confidence interval of that of the RCTs in only 2 of the 19 comparisons. However, in most cases the possibility of a clinically important difference could not be excluded.

A further comparison between effect sizes of randomised and non-randomised studies identified as many as 45 medical interventions for which 240 RCTs and 168 non-randomised studies previously included in meta-analyses of binary data were available for comparisons (Ioannidis *et al.* 2001). Despite a very strong correlation between the synthesised results of both study types ($r=0.75$, $p<0.001$), non-randomised studies tended to show larger treatment effects (28 vs. 11, $p=0.009$). Prospective observational studies tended to fare better than retrospective ones.

A more recent methodological study has been undertaken by Deeks et al. (2003) as part of the UK HTA programme. The work included the generation of non-randomised studies based on re-analysis of data from two large multi-centre RCTs. The researchers re-sampled the data to generate several thousand historically and concurrently controlled non-randomised studies and the same number of RCTs for direct comparisons of results. Distributions of results for each design could thus be compared. The authors concluded that individual non-randomised studies can seriously bias results and mask even moderate treatment effects by reporting significant results in the opposite direction.

A review by Hartz *et al.* (2005) considered that issues of validity are not the only characteristics of a study which can influence its results. In addition, they specified issues of applicability, such as treatment specifics, outcome, and patient characteristics. Unfortunately, these issues were not considered in the previous comparison studies.

In summary, there is some evidence against the assumption that observational studies always arrive at more extreme results than randomised studies. However, given the disparate findings of the available reviews and their own methodological limitations, it is impossible to predict when observational studies might be safe alternatives to RCTs. To that effect, prospective study designs seem to be more promising, as do attention to exclusion criteria and control of confounders.

Despite their systematic approach, the above reviews effectively constitute a series of case studies. Most authors have reported difficulties in identifying studies for inclusion. Thus, the generalisability of their findings cannot be assumed. It is also not possible to draw general conclusions as to whether there are intervention-, patient-, or disease-specific trends in terms of the

validity of non-randomised studies. Further case studies and methodological research will thus be needed.

## 2.7 Comparisons between RCTs and ESDs

The reviews reported in the previous section did not focus on ESDs. I therefore attempted to identify studies which compared RCTs with relevant ESDs. I undertook a Medline search for reviews comparing the findings of RCTs with those from ESDs, using the same search terms as for the review described in section 6.4 ("CHEMICALS AND DRUGS CATEGORY" [MESH] AND (EFFECTIVENESS OR EFFICACY OR OUTCOME OR EFFECT OR EFFECTS) AND (DATABASE OR DATABASES OR DATABASES OR DATABASE OR REGISTER OR REGISTRY OR REGISTERS OR REGISTRIES OR DATASET OR COHORT OR "CLAIMS DATA" OR "SECONDARY ANALYSIS") Limits: English, Human, Core clinical journals).

The search was undertaken on December 2001 and limited to reviews published in high quality English language journals referenced in the Index Medicus. However, this search was abandoned as published abstracts of reviews do not indicate clearly whether ESDs are included in the review. This section therefore only presents relevant studies known to me or encountered during the work on this thesis. This section therefore does not claim to be a systematic review of the subject, but a narrative review that does illustrate different approaches taken by researchers.

Some workers have provided reports of direct comparisons of results from RCTs with those from database analyses. The intentions of their papers vary. Pincus (1993) explored a discrepancy found between one of his earlier analyses of a clinical database (Pincus *et al.* 1992) and a randomised trial (Williams *et al.* 1992). The trial had found no difference between methotrexate, auranofin, and a combination of both in the treatment of rheumatoid arthritis over 48 weeks in days to normalise the creatinine

phosphokinase muscle enzyme, change in muscle strength score, and change in inflammation on the muscle biopsy. The earlier ESD had found that the continuation of methotrexate over a minimum of 5 years was over 50%, that of auranofin less than 10%. When a subset of treatment courses was selected from the database in order to match the inclusion criteria of the trial (only courses representing the first second-line drug used in a patient), no significant differences were seen between the estimated continuation of both treatments after one year. The author concluded that whereas trial and database results concurred when selection criteria and timeframes were matched, the trial failed to highlight clinically important long-term differences (Pincus 1993).

Hlatky and colleagues (1988) have explicitly set out to test whether database analyses could arrive at similar information to RCTs, if the lack of randomisation can be corrected for statistically. The findings of three major RCTs of coronary bypass surgery (Veterans Administration Cooperative Study, European Cooperative Surgery Study, and Coronary Artery Surgery Study (CASS)), were compared with predictions based on statistical models using data from the Duke Cardiovascular Disease Databank. This longitudinal database of patients having undergone cardiac catheterisation is specifically aiming to provide prognostic factors in order to predict the prognosis of individuals. Since direct application of the prognostic model to trial patients was not possible, the authors selected individuals from their Databank who would have been eligible to participate in a particular trial. Their baseline information was then used to predict the results of therapy observed in the trial. In addition, there was an adjustment for the time period of each trial. Predicted 5-year survival was then compared to that observed for the arms of each trial.

The authors report a very close correlation between predicted and observed survivals (Spearman correlation coefficient: 0.73, $p<0.0001$). In five out of six

trial arms, the predictions of the trial results were within the 95% confidence limits of the observed survival rates. It is unclear whether the one inconsistency (medically treated patients of the CASS) was due to a shortcoming of the model or differences between trial and database patient populations. It is suggested that prognostic models may assist in interpreting conflicting RCT results (Hlatky *et al.* 1988). The study considered outcomes after a surgical intervention, and the extent of concordance between predicted and observed effects may be quite different for long-term drug therapies.

There are instances where the conduct of an RCT is closely related to data collection for a patient registry. Feit and colleagues (1999) report for the Bypass Angioplasty Revascularisation Investigation (BARI) the results of a randomised trial (n=1829) of percutaneous transluminal coronary angioplasty (PTCA) and coronary artery bypass graft surgery (CABG). Eligible patients who did not consent to randomisation and for whom treatment was chosen by their physicians were followed-up in an accompanying registry (n=2010). Numerous differences were observed between both studies at baseline. Of 90% revascularised registry patients, 66% were selected for PTCA.

Randomised CABG patients had fewer grafts, fewer lesions, and shorter bypass times than registry patients receiving CABG ($p<0.001$). Patients randomised to PTCA had more significant lesions ($p<0.001$) and were more likely to need emergency CABG than registry patients receiving PTCA ($p=0.027$). Seven-year mortality did not differ significantly between RCT and registry patients (after CABG: 15.6% and 14.2% respectively, unadjusted relative risk: 1.08, $p=0.57$, adjusted relative risk: 0.94, $p=0.66$; after PTCA: 19.1% and 13.9% respectively, unadjusted relative risk: 1.43, $p<0.01$, adjusted relative risk: 1.17, $p=0.16$). The authors did not present a direct comparison of effect sizes of the two studies; rather the emphasis was on the successful selection of patients for PTCA without jeopardising long-term survival (Feit *et al.* 1999).

The CASS of medical and surgical (CABG) interventions for coronary artery disease also operated an RCT and registry (Holloway & Schocken 1988). Chaitman and colleagues (1990) presented 10-year survival data for randomised and registry patients who were eligible but not randomised. No significant differences were found between survival rates of randomised and non-randomised patient groups (medical: 79% and 80% respectively, CABG: 82% and 81% respectively). However, apart from a large proportion of eligible patients who refused random assignment, the registry contained a large number of non-eligible patients. The RCT sample was relatively healthy compared to the actual patient population (Holloway & Schocken 1988).

Feuer and colleagues (1994) have examined whether patients on cisplatin-based combination chemotherapy for advanced testicular cancer and registered in the SEER Programme have experienced the same survival benefit as participants in the Memorial Sloan-Kettering Cancer Centre trials (n=172 and n=133 respectively). Multiple exclusion criteria ensured that both patient groups were comparable in terms of disease staging and prognostic characteristics. Despite most registry patients receiving chemotherapy, the survival of those categorised as having minimal/moderate extent of disease was significantly poorer than that of comparable trial patients (73% and 95% 3-year survival rate; $p<0.001$), for advanced stage patients the difference was less stark (40% and 52% respectively; $p=0.56$). After a range of sensitivity analyses to test the many assumptions involved in defining and comparing the samples, the authors concluded that, rather than the mere availability of effective chemotherapy, other health care delivery factors seemed to have a major impact on patient outcome.

In the area of HIV, several national and international patient databases exist. Phillips and colleagues (1999) have used data from three of these databases to estimate the effectiveness of antiretroviral regimens and compare these

estimates to results of available randomised trials. Three treatment comparisons were identified for which data were available from both sources (trials and databases): (i) zidovudine monotherapy versus combination therapy of two nucleoside analogues; (ii) zidovudine in combination with either didanosine or zalcitabine; and (iii) dual combination versus triple therapy including a protease inhibitor. Cox proportional hazards models were used to analyse data from the three databases for each treatment comparison in an intention-to-treat analysis, adjusting for baseline differences and confounders.

The adjusted relative risk estimates for progression to AIDS or death were compared with corresponding trial results. For comparisons (i) [zidovudine monotherapy versus combination therapy of two nucleoside analogues] and (ii) [zidovudine in combination with either didanosine or zalcitabine], the estimates from the databases were similar to those resulting from the trials (comparison (i): between 0.61-0.84 compared with 0.57-0.63 respectively; comparison (ii): between 0.81-1.01 compared with 0.77-0.92 respectively). In comparison (iii) [dual combination versus triple therapy including a protease inhibitor], one of the cohorts contradicted the trial result (relative risk: 1.20 (95%CI: 1-01-1.44) compared to 0.50 (95%CI: 0.33-0.76) respectively). The authors conclude that biases were present albeit small in eight out of nine comparisons. Given the inability of large-scale trials to keep up with the speed of development of HIV therapies, the consideration of a combination of several database analyses seems very useful, if prognostic factors are balanced.

In summary, these case study comparisons of RCTs and published ESDs show a similar picture to that presented by the reviews in section 2.6: patient selection appears to be an important issue in terms of comparability, as are other healthcare related factors which may not always be well described or controlled. It is difficult to search for such studies comprehensively, and hence to draw any methodological conclusions from these selected examples.

## 2.8 Summary

Randomisation is the preferred method to assure internal validity of studies evaluating treatment effects. However, RCTs often only achieve poor external validity. Moreover, the trial environment and randomisation per se may introduce different biases and may therefore limit a realistic evaluation of treatment effectiveness. Alternative randomised study designs have been proposed to address specific criticisms of RCTs, but experience with them is still limited.

Systematic comparisons of randomised and non-randomised studies have been inconclusive, but it would appear that there is no systematic distortion of effect estimates stemming from non-randomised studies. It has been suggested that effect estimates from both study types are generally similar if baseline differences are adjusted for in the non-randomised studies, and the same selection criteria are used (Britton *et al.* 1998). None of the systematic comparisons focussed on ESDs specifically. Similarly, several reviews have assessed the potential contributions of database research to health technology assessment, but their focus is also not on ESDs, let alone ESDs of drug therapies.

ESDs of treatment effectiveness seem to offer potentially comparatively large-scale assessments with long follow-ups of treatment use and associated outcomes in naturalistic settings. Such studies have been much debated and criticised, particularly for their proneness to confounding by indication and poor internal validity. Whereas guidance on the conduct of such studies is increasingly becoming available, no systematic assessment or review of them has yet been undertaken. There are limited examples and limited experience with comparisons of non-randomised and randomised studies, particularly

ESDs, and reviews or single comparative studies do not allow any firm conclusions on the validity or quality of ESDs in general.

The following three chapters describe a case study of an ESD (the Dornase Alfa Case Study), with particular focus on the methodological challenges arising in the process. This is followed by a comprehensive review of similar published studies and their methodological features relating to internal and external validity, as well as an attempt at comparing their results with those from RCTs.

# 3 Dornase Alfa Case Study: Preparatory Issues

## 3.1 Introduction

This and the subsequent two chapters report an attempt to undertake an ESD. The intention of this chapter is to describe the design and preparatory phases of the study and the challenges encountered there. The subsequent chapter illustrates in detail further methodological challenges encountered in examining the data source and the quality of its data. These chapters thereby demonstrate some of the difficulties and weaknesses of ESDs. The study itself has been reported previously to the funding body (the NHS Executive North West) by the principal investigator (Dr A Haycox), and the third chapter dedicated to the Dornase Alfa Case Study (Chapter 5) provides an overview of the descriptive analysis undertaken by an analyst at the time.

Cystic fibrosis (CF) is an inherited disease, affecting an estimated one in 2,500 newborn babies in the UK (Dodge *et al.* 1997). The genetic defect is associated with abnormally high absorption of sodium between the cells, resulting in thickened, dehydrated secretions. The most serious manifestation is respiratory disease. Patients have abundant viscous sputum, which can become infected by bacteria. Inflammation of the airways contributes to progressive lung damage. The median life expectancy of people with CF has increased significantly over the past decades due to comprehensive treatment regimens. In 1991, the median life expectancy of a child born with CF has been estimated to be around 40 years (Elborn *et al.* 1991).

Recombinant human deoxyribonuclease (rhDNase) called dornase alfa is a genetically engineered version of a human enzyme which cleaves extra-cellular deoxyribonuclease (DNA). It is marketed under the name *Pulmozyme®*. The drug is administered by inhalation and cleaves the DNA in the sputum, thereby reducing its viscoelasticity. Clearing the lungs of the viscous sputum (usually through physiotherapy) is one of the most important

interventions in CF, and dornase alfa is intended to facilitate that process. The drug has been licensed since 1994 and the current UK licence is for use in CF patients who are over five years of age with a forced vital capacity (FVC) of greater than 40% of predicted, to improve pulmonary function.

The annual treatment costs in the UK for continuous dornase alfa therapy are currently approximately £6,800 per patient. There is evidence for the short-term effectiveness of the drug, but the long-term effectiveness is still under question. It was clear that a long-term RCT was not feasible due to most existing patients no longer being naïve to the drug. Any new potential RCT would have had to recruit from a small number of newly diagnosed patients meeting certain eligibility criteria. Also, the loss of equipoise would have spoken against another RCT. The NHS Executive North West therefore provided funding for an observational study in 1999, to elicit evidence on the long-term cost-effectiveness of dornase alfa.

This and the subsequent chapter present the preparation phase and the methodological challenges encountered in the process. Subject-specific information, such as the complete literature review and more detailed descriptive Dornase Alfa Case Study results are presented in Appendix A to not deter from the methodological focus. This chapter describes how the original study design had to be adapted sequentially due to practical difficulties with the planned approaches. One of the main issues here was data protection, and this topic is covered in a separate section.

### 3.1.1 Summary of the literature review

Readers unfamiliar with the nature and epidemiology of CF, or the literature on the effectiveness of dornase alfa, may wish to refer to Appendix A for a detailed literature review undertaken for the Dornase Alfa Case Study. Below,

the conclusions relevant to the rationale for an observational study approach are reproduced:

- There were indications that benefits of dornase alfa in CF in terms of a reduced lung function decline and a reduced risk of respiratory exacerbations may continue over the long term. Randomised controlled trial evidence of the efficacy of dornase alfa was available for a follow-up period of 2 years. In the longest (two-year) RCT, the treatment group exhibited a significantly reduced lung function decline vis-à-vis the control group. However, after one year, this difference was not significant (Quan *et al.* 2001).
- Longer-term RCTs do not exist and are very unlikely to be possible in future. Observational studies of more than two years follow-up are rare; they are so far of relatively small scale and frequently without comparison groups. The authors of a relatively recent four-year case-control study recommended further long-term studies involving larger cohorts (Shah *et al.* 2001).
- Considerable variations in the nature and extent of response between and probably also within individuals as well as practice variations between different CF centres add to the difficulties in evaluating the effectiveness of dornase alfa.

### 3.1.2 Aim and objectives

The aim of the original Dornase Alfa Case Study was to evaluate whether long-term use of dornase alfa was associated with a delay in disease progression in CF patients and hence an extension in their survival, and to assess whether statistically significant changes in lung function measures demonstrated in the trials would translate into clinically relevant outcomes such as exacerbations, or disability / quality of life.

The objectives of the study were to:

1. undertake a long-term comparative analysis of homogenous groups of dornase alfa users employing a large scale longitudinal database;
2. analyse the mortality experience of the different dornase alfa user groups, as far as is possible within the sample;
3. analyse the rate of respiratory exacerbations experienced by patients from each group;
4. analyse the long-term decline in lung function experienced by each dornase alfa user group;
5. analyse the long-term anthropometric changes experienced by each dornase alfa user group.

As will be demonstrated, the analysis of the data was limited for a variety of reasons. Therefore, the intention of this and subsequent chapters describing the Dornase Alfa Case Study focuses on learning methodological lessons from the detailed process of preparing and conducting the study.

## 3.2 Proposed and final study designs

The research question of the protocol approved for funding by the NHS Executive North West was *"to undertake a rigorous naturalistic multi-centre trial to assess the impact of DNase on disease progression, quality of life and resource consumption in patients with cystic fibrosis"*. It was intended to exploit the difference in treatment protocols between centres, whereby some used dornase alfa, and others did not. Thus retrospective data collection should have first identified matched pairs of patients from seven Northern English CF centres (three adult and four children's centres) - patients on and off dornase alfa. A retrospective study phase should have reviewed available data and thus permitted sample size calculation for a prospective study phase. In a 1-year prospective phase data on relevant outcomes, specifically long-term trends in lung function, and self-rated quality of life, would have been collected for comparative analysis of patients on and off the drug.

The project steering group included the principal investigator and all co-applicants for the original proposal funded by the Department of Health, North West Regional Office.[1] Prior to commencing the study, early discussions with the steering group revealed that the proposed prospective phase of the study in particular was considered not feasible in practice. Instead, a solely retrospective record review was suggested, covering a longer time period of

[1] Prof Deborah Ashby — Bart's and the London Queen Mary's School of Medicine and Dentistry, University of London
Dr Steve Conway — St James' & Seacroft University Hospitals, Leeds
Prof Tim David — Booth Hall Children's Hospital, Manchester
Dr Garry Hambleton — Royal Manchester Children's Hospital
Dr Alan Haycox — University of Liverpool
Dr David Heaf — Royal Liverpool Children's NHS Trust
Prof Rosalind Smyth — Professor of Paediatric Medicine, University of Liverpool
Dr Maurice Super — Royal Manchester Children's Hospital
Prof Tom Walley — University of Liverpool
Dr Martin Walshaw — Cardio-Thoracic Centre, Liverpool
Prof Kevin Webb — Wythenshawe Hospital, Manchester

10 years, and comparing the experiences of patients "on" and "off" dornase alfa. Data from this period was to be collected from medical records and existing information systems in the participating seven CF centres, and thus avoided direct data collection involving patients.

A pilot phase sought to establish the availability of suitable data in the centres and to estimate the time necessary for data extraction. A question pro forma was developed to interview the centre clinicians about data availability and accessibility of key variables such as demographic data, lung function measurements, anthropometric measurements, dornase alfa use, and exacerbations. Clinicians and operational staff in the centres were interviewed by telephone to elicit this information.

The retrospective data collection was piloted in one centre which was considered to have relatively well-summarised and accessible paper-based records going back several years. Even there, records did not reach far enough back, and what was more important, the manual data extraction per patient took far longer than the estimated time available for the retrospective study phase. Whereas there was one other centre which had computerised data, this was rather limited in terms of available variables. In other centres, actual records would have had to be reviewed, which would have multiplied the time necessary.

The interviews and pilot work further identified that (1) data availability for key variables covering the period in question varied considerably between the units, and, given that I would have had to access and retrieve data from patient records, (2) further significant delays would have been inevitable because of the need for getting individual written consent from patients, parents, or guardians.

The Steering Group therefore identified another option: members considered that all centres involved in the study had over a number of years contributed anonymised data on their patients to a Europe-wide database of CF patients, the European Registry of Cystic Fibrosis (ERCF). Many of the data items that were originally to be retrieved manually from patients' records were collected in that database. The principal investigator therefore was tasked with approaching the chair of the ERCF Advisory Committee to seek access to the data.

The Europe-wide Advisory Committee for the ERCF was keen for researchers to use the data (not least as this was hoped to support the case for applying for public funding to continue the registry), but had not yet been approached by any outside institutions wanting to do so. Consequently, no formal process for gaining access to the data existed. The ERCF Advisory Committee controlled access to the data and considered the individual CF centres and patients to be the owners of the ERCF data. The Committee had to first explore and agree their approach for releasing the data. The process between the first application to the Advisory Committee for access to the ERCF data and the eventual receipt of the dataset took 6 months.

Agreements had to be reached on the proposed use of the dataset, as well as the potential restrictions on any future publications resulting from the analysis. However, a main concern of those developing the process for granting the research team access to the data appeared to be the confidentiality of centres and clinicians, as much as patients. Between the Advisory Committee and the company managing the ERCF (*NSPM*), it was decided that each CF centre should give written permission for its data to be made available for the current research study. In addition, principal investigators in the seven participating units had to give their written permission for the research team to receive the code which identified their

respective centres (rather than patients). The process from my appointment to this study to the receipt of the first copy of the ERCF took one year.

Clinicians on the local Steering Group had expressed reservations about the quality and completeness of data in the ERCF, and therefore detailed data verification was planned against information existing in the centres. This meant that individual patients' data had to be decoded and checked. Quintiles, the company holding the ERCF data, was instructed to introduce an additional layer of coding for extra protection, so that only centres themselves were able to identify patients from the dataset. This meant that CF centres involved in the data verification exercise had to go through two steps of decoding to identify patients from the dataset.

However, the data verification also encountered data protection issues arising from new legislation at the time, as the Data Protection Act had only just been published and was not consistently understood and interpreted by relevant decision-makers in participating centres. The following section considers the impact of this new legislation on research using patient records. It describes how I was forced to navigate and interpret the then very unclear and conflicting available guidance, at the same time as the research approach kept being adapted and therefore raised different data protection issues each time.

## 3.3 Data protection and confidentiality issues

### 3.3.1 Introduction

As described above, the protocol for the research project underwent considerable changes. Each of them posed questions regarding applicable data protection and confidentiality rules. It is important to stress that this work was undertaken at a time of major changes in NHS governance, particularly research governance, around the year 2000, and against the background of inconsistent and unclear policy and guidance from various bodies. More guidance exists now, and this is referred to in the Discussion (Chapter 8).

This section describes in detail the data protection and confidentiality issues faced at the planning stage of the research project. Firstly, the section reports on the process of gaining approval by the Research Ethics Committees. Thereafter, the changing research approach and resulting data protection issues are mapped out for clarity, and the reader taken on a chronological journey of the decision-making processes involved, accompanied by summaries of the then existing guidance.

### 3.3.2 Research ethics committee approval

Before work on the research project began, approval for the initial protocol was sought and received from the North West Multi-Centre Research Ethics Committee (MREC), as well as the relevant Local Research Ethics Committees (LRECs) for seven participating centres. At the time, the latter worked independently of the MREC and could in theory reject a study approved by the MREC. In this case, only minor conditions were imposed, e.g. to use Trust-

headed paper for correspondence with patients. A later protocol alteration was similarly approved by the MREC, and the required written notification of these alterations was sent to the relevant LRECs.

Research ethics committee approval is not synonymous with the fulfilment of data protection and confidentiality laws, but constitutes considered advice on the observance of ethical principles in research proposals. As soon as the first changes to the protocol were discussed, the MREC's advice was sought on the need for consent of patients to permit the collection of data from their medical records by myself. The MREC advised us to check the proposal with data protection officers of the responsible hospital Trusts (now this is national guidance). This was the beginning of a lengthy exploratory journey through then current data protection legislation and guidance. The following section illustrates this alongside the changes in the research approach.

### 3.3.3 Changes in the research approach and resulting data protection issues

The initial research protocol envisaged the collection of original data from patients (e.g. quality of life measurements) as well as the access to patient record systems by myself as a University-employee to retrieve routinely collected clinical data. The second protocol involved only the retrieval of routine data.

The final protocol meant that routine data submitted to a clinical database, the Epidemiologic Registry of CF (ERCF), were to be used (see section 4.2.1). This database had been set up in 1994 by F. Hoffmann-La Roche across several European countries as a multi-centre, longitudinal, follow-up project of CF patients receiving routine care. The intention was to collect data on all patients of participating CF units and analyse progression of lung disease by pulmonary function and infection rates (Anonymous 1994). The data were

anonymised, insofar as the CF centres reporting data relating to individual patients were able to decode the data and identify their patients. For the planned extensive data verification process as part of this research, however, this decoding was unavoidable.

The table below illustrates the relevant changes in research approach and the data protection issues arising from each. It is obvious that the issue of "consent" (to accessing records or coded data) played a key role. The requirement for this was not clearly defined at the time, and subsequent questions arose from the available guidance, e.g. under which conditions might consent be implied, rather than explicit (e.g. written) consent be necessary, and under which conditions might a requirement for consent be waived? Another key question was this: whose role and responsibility should it be to decide upon the answers?

**Table 3.1: Changes in research approach and data protection issues**

| Research approach | Data protection issues |
|---|---|
| **First protocol:** | |
| (1) Data collection from patient records, plus | (1) Is consent required for access to records by a university-employee? |
| (2) Prospective data collection from patients and records | (2) Clear that consent was required for prospective phase. |
| **Second protocol:** | |
| (1) Longer retrospective data collection from patient records | As (1) above |
| **Final protocol:** | |
| (1) Use of ERCF database for analysis | (1) Is consent required given that this is "coded" data, i.e. the decoding key still exists? |
| (2) Decoding of individual data in centres for data verification exercise | (2) Only staff ordinarily able to access patients' data should be involved in this exercise |

This story cannot be told without equipping the reader with the basic concepts involved, and introducing him to the then available guidance and resulting issues. In doing so, I re-trace my steps in seeking clarification from available guidance, as I was increasingly becoming aware of the complexities involved. However, as will become more obvious, the guidance documents were inconsistent and confused the issues. I finally published part of our experience in a paper in the British Medical Journal - possibly the first having illustrated the implications of the then current state of the law for research involving patient records, and hence a major contribution to the debate at the time (Strobl et al. 2000). It is worth pointing out that further developments and debates have taken place since, and an update can be found in Chapter 8.

Before describing the sources of law and guidance at the time, I will introduce the reader to definitions of relevant concepts (and their lacking clarity): Table 3.2 illustrates well the confusion already arising from merely attempting to clarify the concepts involved in answering the question of consent for accessing records.

**Table 3.2: Definitions and their inconsistencies at the time**

| |
|---|
| **Personal data**<br>The EU Data Protection Directive 95/46/EC defines personal data as "any information relating to an identified or identifiable natural person" (European Parliament and Council 1995 p.42). A person may be identifiable directly or indirectly, also through the use of identification numbers or reference to specific items of information. The UK Data Protection Act 1998 limits this to living individuals. |
| **Data processing**<br>According to the Data Protection Act 1998, data processing means obtaining, recording, or holding data and includes any operation performed upon the data, whether automated or not. Disclosure is a data processing operation. |
| **Data controller**<br>Legal person or agency who determines the purpose and means of processing (European Parliament and Council 1995). The data processor may be different from the data controller. |
| **Common Law Duty of Confidentiality**<br>The British Medical Association defines confidentiality as "the principle of keeping secure and secret from others, information given by or about an individual in the course of a professional relationship" (Romano-Critchley & Sommerville 1999). This legal duty applies to information entrusted to someone in confidence. The duty of confidentiality applies independently of the UK Data Protection Act. |
| **Consent for data processing**<br>The EU Directive 95/46/EC defines consent as "any freely given specific and informed indication of his wishes by which the data subject signifies his agreement to personal data relating to him being processed" (European Parliament and Council 1995 p.43). Rather unhelpfully, the UK Data Protection Act 1998 does not define consent at all.<br><br>However, the definition given by the Directive leaves uncertain the nature of the required consent. For example, there is uncertainty about whether consent can be implied if the patient knows that data will be shared but does not express any objection. Also, it remains unclear what constitutes information for "informed" consent. Seemingly unrelated to the issue of consent, the Directive specifies information to be given to the data subject; this may be insufficient as a basis for his/her decision to permit their data to be processed in the first place.<br><br>According to the Directive, the "unambiguous" consent of the data subject is one of several criteria for making data processing legitimate; but consent can be waived under certain conditions, such as the vital interest of the data subject, or if processing is in the public interest.<br><br>In relation to health data, the Directive - as well as the UK Data Protection Act 1998 - requires "explicit" consent of the data subject to the processing of such data, unless other conditions are met. Among these conditions are purposes of preventive medicine, medical diagnosis, provision of care or treatment, or health-care service management, as long as a health professional or person subject to obligations of secrecy processes the data. EU member states are permitted to lay down additional exemptions for reasons of "substantial" public interest, either by law or decision of the supervisory authority (in the UK, the Data Protection Commissioner, see below). |

| |
|---|
| **Anonymisation**<br>The EU Data Protection Directive concerns data which may identify individuals directly or indirectly. Many read this as including coded data, not least the UK Data Protection Commissioner, who uses the term "pseudonymised" in this case. Some guidance documents interpret anonymisation differently, however. For example, the Royal College of Physicians Committee on Ethical Issues in Medicine (1999) recommends early anonymisation of data for research using archived patient data. However, the authors must have been aware that longitudinal databases cannot do away with codes, otherwise updating of registers, or data verification would not be possible. Even where guidelines talk about a complete separation of the clinical or administrative information from anything which may identify an individual, it still remains unclear whether data can be called anonymous if the decoding information remains in existence but inaccessible to e.g. a researcher. |
| **Public interest**<br>The EU Directive 95/46/EC permits processing of personal data in the public interest (even without consent). It is noteworthy that in the context of health data, the Directive uses the expression "substantial public interest". However, the concept of public interest (or for that matter, substantial public interest) is not defined. The BMA guidance (Romano-Critchley & Sommerville 1999) gives much attention to this issue. |
| **Audit or research**<br>In the UK, audit does not require consent or ethical approval (Royal College of Physicians Committee on Ethical Issues in Medicine 1999), albeit that this view is not held unanimously, and that there is still some confusion as to the dividing line between audit and research (Wilson et al. 1999). The UK Department of Health acknowledges that at times the dividing line is blurred.<br><br>Whereas some hold that audit is part of the healthcare process and therefore does not need consent (R&D Office 2002), others view patients' consent for audit as necessary unless their data are effectively anonymised (GMC 1998). |

Various professional and health sector bodies have addressed questions of consent, anonymisation of data for research, and access to medical notes for research purposes (rather than audit), and I have already drawn on existing guidance documents in defining the concepts above. At the time of protocol development for the Dornase Alfa Case Study project, some of these documents were in the process of being updated, some had already been published and met with criticism, not least for pre-empting others (Romano-Critchley & Sommerville 1999).

With few exceptions, broad debate about the implications of the then new Data Protection Act 1998 was lacking, particularly in the context of

epidemiological research using patients' records. Thus, there was a dearth of up-to-date and clear policy guidance on which to base research protocols. This was not helped by the fact that the interpretation of the Act as well as of the common law duty of confidentiality were subject to debate, and no case law existed which might have clarified at least some of the interpretation.

This section summarises the relevant legislation and guidance available at the time of planning the research project, with an emphasis on the issues raised by the project. It is useful to revisit the relative hierarchical positions of the guidance and legal instruments. A Directive in European Union law is an instrument of primary legislation; it is binding to all EU member states which have to implement it in their national law. In the UK, constitutional law consists of statute law and case law, with the latter constituting judicial precedents as judges in the courts interpret statute law (Carter 2002). Public General Acts (such as the Data Protection Act 1998, or the Human Rights Act 1998) and Local and Personal Acts are the two types of primary UK legislation.

Statutory instruments (SIs) are regulations made under the authority of an act and thus constitute secondary legislation. In contrast to the planning phases of the research study, there are now several statutory instruments relating to the Data Protection Act 1998 clarifying particular aspects of the Act. For example, statutory instruments have expanded on the access of data subjects to their own data, including an SI referring to health data (Anonymous 2000b), or on conditions under which sensitive personal data may be processed (Anonymous 2000b). However, no SI seems to clarify in particular any of the issues faced in the research project described in this thesis.

In the case of the Data Protection Act 1998 in particular, the views and guidance expressed by the UK's Information Commissioner (formerly Data Protection Commissioner) interpreting the Act can be viewed as authoritative. The Commissioner has published two significant documents which became

available only after the critical decision-making phase in the Dornase Alfa Case Study project (Information Commissioner 2001; Information Commissioner 2002).

Guidelines and codes of professional and public bodies do not constitute any legal instrument, but they may be taken into consideration in a court of law. One SI (Anonymous 2000a) has accepted certain codes published by agencies such as the British Broadcasting Corporation, Press Complaints Commission, or Independent Television Commission as relevant for interpreting the Act. Compliance with guidelines of a professional body may be looked upon favourably by a court of law, but it does not guarantee legality of conduct.

**European Directive 95/46/EC**

The *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data* (European Parliament and Council 1995) stresses the Member States' responsibility to protect the right of natural persons to privacy with respect to the processing of personal data, whether automated or not.

A guiding principle of the Directive is that personal data must be processed "fairly and lawfully". Neither the Directive, nor the Data Protection Act 1998 made clear the meaning of this phrase, and there was room for different opinions on its interpretation and thus implementation. In the context of health data, the phrase was generally interpreted as complying at least with the common law duty of confidentiality.

Whereas data must not be processed in a way incompatible with explicit and specified purposes, further processing for statistical and scientific purposes is permissible.

**Data Protection Act 1998**

This brings into UK law the European Directive 95/46/EC on the processing of personal data. The Act came into effect on 1 March 2000, and in comparison with the 1984 Data Protection Act (which it replaces) is concerned with both records on paper and records held on computers. The Act does not concern personal data relating to dead individuals, or anonymous data (which is no longer considered identifiable and thus not personal). The Act is based on eight principles (Table 3.3), the first of which stipulates that "*personal data shall be processed fairly and lawfully.*" Schedules 2 and 3 of the Act set out conditions applying for the purposes of this First Principle.

**Table 3.3: Eight principles of the Data Protection Act 1998**

| | |
|---|---|
| (1) | Personal data shall be processed fairly and lawfully and, in particular, shall not be processed unless -<br>a. At least one of the conditions in Schedule 2 is met, and<br>b. In the case of sensitive personal data, at least one of the conditions in Schedule 3 is also met. |
| (2) | Personal data shall be obtained only for one or more specified and lawful purposes, and shall not be further processed in any manner incompatible with that purpose or those purposes. |
| (3) | Personal data shall be adequate, relevant, and not excessive in relation to the purpose or purposes for which they are processed. |
| (4) | Personal data shall be accurate and, where necessary, kept up to date. |
| (5) | Personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes. |
| (6) | Personal data shall be processed in accordance with the rights of data subjects under this Act. |
| (7) | Appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data. |
| (8) | Personal data shall not be transferred to a country or territory outside the European Economic Area unless that country or territory ensures an adequate level of protection for the rights and freedoms of data subjects in relation to the processing of personal data. |

Schedule 2 applies to the processing of any personal data and stipulates that data subjects have to have given "consent" to processing (which is not further defined), or the processing has to be necessary for a variety of stated reasons.

Among these are the "exercise of any other functions of a public nature exercised in the public interest by any person", and circumstances that the Secretary of State may specify as purposes of legitimate interest.

Schedule 3 applies to the processing of sensitive personal data (in addition to Schedule 2), which includes health data. For the processing of such data, "explicit consent" is required. Again several alternative conditions are listed. One of these relates to processing being necessary for medical purposes - which explicitly include medical research -, as long as a health professional or someone with a duty of confidentiality undertakes them. It is important to remember, that the First Principle of fair and lawful processing needs to still be met in medical research, but that this as well as confidentiality were poorly defined.

In Section 33, the Act makes an exemption for research with regard to the Second Principle. Accordingly, the further processing of personal data for research is not incompatible with the purposes for which they were obtained, provided the processing does not get used with respect to individuals or is likely to cause distress or damage to a data subject. This exemption comes into play where health professionals use for research the data they have collected "fairly and lawfully" as part of their professional relationship with patients.

**Human Rights Act (1998)**

This Act formally incorporates into UK law the European Convention of Human Rights. This guarantees the right to respect of privacy. The Act enables people to enforce their Convention rights in the UK and makes it illegal for public authorities to infringe these rights. It was thought that the impact of this Act on current practice would be small, as current ethical standards were seen as likely to be compliant with the Act (BMA 2000).

**Case law**

Ultimately, the legality of using confidential data in research can only be determined on a case by case basis by the courts (Strobl *et al.* 2000). However, there are few court rulings to guide the interpretation of the common law of confidentiality (see next section).

**Department of Health guidance**

The Department of Health did acknowledge that there were conflicting legal views on applying the duty of confidentiality and was at the time trying to interpret it for the health sector (NHS Executive 1999c). A previous 1996 Health Service Guidance document (HSG(96)18) is quoted which holds that consent to sharing information for NHS purposes can be implied if a patient has been informed and does not object.

In *For the Record* (NHS Executive 1999b), it is noted that the common law duty of confidentiality can be overridden by public interest. The document goes on to say that if data are anonymised ("so that individual patients cannot be identified"), it can be used for epidemiological research (presumably without consent).

The Department of Health was involved in a court case brought by Source Informatics Inc. who sought to obtain and sell anonymous information from pharmacies (Anonymous 1999). The first ruling in the case raised the question of whether even the processing of fully anonymous data was in breach of confidentiality (Anonymous 1999). However, an appeal against this was upheld (Anonymous 2000d).

Thus at the time of the research proposal development, topical Department of Health guidance was lacking, albeit that it had been announced and was imminent. In the meantime, UK professional and research bodies had

developed their own (conflicting) guidance documents, which are summarised in Table 3.4. In summary, at the protocol development stage of the Dornase Alfa Case Study, there were significant gaps and uncertainties across the legislative and guidance documents in the definitions of key concepts, such as what constituted "fair and lawful" processing, the duty of confidentiality, consent, as well as "public interest". Professional bodies had issued guidance documents, which could be seen as incongruent with the Data Protection Act. It may be suspected that some incongruence was a deliberate challenge, as the medical and research community was becoming aware of the potential threats to the epidemiological research posed by the legislation. Only the resulting confusion was certain.

**Table 3.4: Guidance from UK professional and research organisations**

| |
|---|
| ***General Medical Council (GMC)***<br>The GMC guidance *Seeking Patients' Consent: the Ethical Considerations* (GMC 1998) referred partly to research. The GMC demanded that doctors seek written consent for research from patients, but did not specifically refer to the use of anonymous data for research. |
| ***British Medical Association (BMA)***<br>The BMA had published one of the most recent and most detailed guidance documents available at the time (Romano-Critchley & Sommerville 1999). The guidance held that passing patient information from one health professional to another in the "NHS family" who is not involved in the care of the patient concerned constitutes a breach of confidentiality. The guidance made surprisingly few references to the Data Protection Act. Instead, the BMA took a clear stance on anonymisation, stating that the processing of fully anonymised data did not require consent, but advised clinicians to seek legal advice given the uncertainties.<br><br>Most critically, perhaps, this guideline defines *personal health information* as one from which a person can be identified. The guideline goes on to say that *"Information which does not permit the recipient to identify an individual is not contentious. Coded, encrypted, aggregated, or anonymised data can easily be used effectively for many health service or research purposes instead of identifiable data. ... Use of minimal data identifying the patient's electoral ward, sex and year of birth is acceptable for administrative or research purposes"* (Romano-Critchley & Sommerville 1999).<br><br>Anonymised information is defined as information which does not identify the data subject, either directly or indirectly. The text following this definition in the guideline, however, seems to imply that what matters is that the user of the data is unable to identify the data subject. This and the paragraph above indicate that the BMA's view of anonymisation differs from that of the Data Protection Commissioner, who holds that for truly anonymous data no decoding key should exist anywhere.<br><br>Lastly, the guidance seems to imply that patient records can be used for research (presumably by the treating clinicians), but health professionals must make a "reasonable effort" to inform patients of this possible use of their data and give them the opportunity to object. |
| ***Medical Research Council (MRC)***<br>The MRC had published a draft guideline for consultation on *Personal Information in Medical Research* on 17 September 1999. It stated that *"Most international and national ethical codes do not require explicit, informed, consent for research based only on records that will not directly affect the individual"* (MRC 1999). The guideline was based on legal advice to the MRC and emphasized the responsibility of individual clinicians for safeguarding confidentiality, but also acknowledged that legality of any processing of personal data would have to be determined on a case-by-case basis.<br><br>The guideline put great emphasis on informing patients routinely about possible use of records for research in general; patients should be given the opportunity to object to the use of their data for research. It was thought that this would *"go a long way towards meeting the legal principle of "fair processing"."* |

| |
|---|
| Regarding anonymisation, the guideline stated that personal information must be encoded *or* anonymised as far and as early as possible. Most importantly, in the MRC's view, if consent was impractical, it was permissible to disclose personal information for research under certain conditions. These conditions are related to safe processing, except one: the importance of the study justifying the disclosure. Research Ethics Committees were supposed to judge whether this was the case. |
| ***United Kingdom Central Council for Nursing, Midwifery and Health Visiting (UKCC) (now: Nursing and Midwifery Council, NMC)***<br>The UKCC's guideline on records and record keeping stated that the principles of confidentiality apply to the use of records for research, and the right of the patient to refuse access should be respected (UKCC 1998). Whereas it clearly stated that explicit consent was needed for disclosing patient data (unless disclosure is required by law or is necessary in the public interest), the UKCC was less clear on what it meant by disclosure. No explicit consent was mentioned where the use of records for research were discussed (UKCC 1996). |
| ***The Royal College of Physicians' Committee on Ethical Issues in Medicine*** (1999) had published recommendations relating to research based on archived information. The Committee expressed concerns over the severe restrictions the Data Protection Act placed upon non-intrusive record-based research and held the view that such research should be possible without the express consent of data subjects. A condition for such research should be the anonymisation of the data at the earliest opportunity and to a minimum level of precluding identification of individuals from the output of the research.<br><br>Significantly, the Committee specifically recommended the Department of Health not to subscribe to legislation which could impede such research. This was at the time when the Department of Health was preparing its own guidance on the Data Protection Act and had just lost its appeal in the Source Informatics case (Anonymous 2000d). |

The above presented results of my enquiries into existing concepts, definitions and guidance meant that there would be no agreed authoritative answer. Before I go on to describing the next steps in decision-making, it is worth considering the different potential participants involved in this process. These are described in Table 3.5.

**Table 3.5: Participants in the Decision-Making Process**

| |
|---|
| **Caldicott Guardians**<br>In 1997 the Caldicott Committee reported on its review of information that identifies NHS patients (Caldicott Committee 1997). In keeping with the report's main recommendations each health authority, Trust, and primary care group in the United Kingdom appointed a "Caldicott Guardian." A key responsibility of these Guardians was to agree and review local protocols for the protection and use of identifiable information obtained from patients, including protocols governing disclosure across organisational boundaries (NHS Executive 1999a). |
| **Data Protection Commissioner (now: Information Commissioner)**<br>This office exists for the purposes of the Data Protection Act 1998 (the role had been established in the previous Data Protection Act 1984 as Data Protection Registrar). The role of the Commissioner is to promote the observance of the Act. He or she may also assess any processing operation for good practice. |
| **Data Protection Officers**<br>Data controllers, such as NHS Trusts, or universities, have appointed named individuals who lead on their organisation's endeavours to comply with the 1998 Data Protection Act. They ensure the formal notification of all uses of personal data to the Data Protection Commissioner, and act as internal advisors and facilitators. |
| **Clinicians**<br>Clinicians remain accountable for the confidential information they have received from their patients, regardless of authorisations by Research Ethics Committees or even Health Authorities (MRC 1999). |
| **Research Ethics Committees**<br>The NHS has a system of Local and Multi-Centre Research Ethics Committees to which any proposal for research within the NHS has to be submitted for consideration. Department of Health guidance was very clear on the need for approval by a Research Ethics Committee for any research using patient records (NHS Executive 1999b). The Royal College of Physicians' Committee on Ethical Issues in Medicine (1999) seemed less agreed on this issue.<br><br>Research Ethics Committees, however, cannot be asked to interpret the law on behalf of applicants or data controllers involved, and ethics committee approval of a research project is not necessarily an indication that it complies with the law. Although committee approval is necessary for research within the NHS, research ethics committee decisions have no legal standing (MRC 1999). Committees simply advise on whether a proposed project meets recognised ethical standards. |
| **Data controllers of the database**<br>The ERCF arguably held pseudonymised data, whereby the decoding key was held by the respective CF centres. Apart from CF centres who contributed the data, several organisations were involved: funding came from a pharmaceutical company (Roche |

| |
|---|
| Pharmaceuticals), the management of the Registry was ultimately entrusted to a Swiss company (NSPM, Lucerne, Switzerland), whereas the data handling was contracted to Quintiles (Frankfurt, Germany), a pharmaceutical services company. The Steering Group was made up of an international group of clinicians, but individual clinicians in the participating centres had a key role in deciding over access to data from their own centres. |
| **Patients**<br>It is ultimately the patients' data which is sought to be protected. However, patients had little part in the discussion, and despite publications of the issues in high-profile journals, there was never any real public debate on data protection and epidemiological research. There are now surveys which suggest that patients are relaxed about such use of their data (Barrett *et al.* 2006), however, others show that the public clearly want to be asked for their permission (Clayton 2008). |

Therefore, at the outset of the study, and following MREC approval and advice, I asked the Data Protection Officers of participating hospital Trusts for their advice on the need for consent from patients, given that I - a University employee - sought to access patient records for retrospective data collection. The responses to this request are shown in the Table 3.6. The Trusts' decisions varied considerably and usually involved discussions and consultation between Data Protection Officers, Caldicott Guardians, and at times executive directors; consequently, this led to delays.

**Table 3.6: Trusts' decisions on whether patients needed to give explicit consent**

| | Decision | Time to decision |
|---|---|---|
| Trust 1 | This Trust decided that the researcher could have access to patients' records without explicit consent from patients as long as no identifiable information was removed from the hospital (for example, the researcher could extract information from records and retain it in coded form but the key for decoding would be kept at the hospital). | < 3 weeks |
| Trust 2 | The Caldicott guardian decided that consent from patients was required. This decision was later revised after the Trust sought legal advice, and the researcher was then permitted to have access to patients' records because the Data Protection Act 1998 only came into force after the start of the study (1 March 2000). | 4-5 months |
| Trust 3 | The data protection officer and the Caldicott guardian advised the researcher to obtain explicit consent from patients because the researcher was not a staff member of the Trust and no explicit consent exists from patients to permit the use of their data for research (for example, no agreements are signed by patients when they are first seen). | 6 weeks |
| Trust 4 | The data protection officer immediately decided that the proposed study required explicit consent from patients since only staff with a duty of care to the patient are permitted to have access to that patient's medical records, and, unlike audit, research is not seen as part of the healthcare process. | Immediate |
| Trust 5 | The data protection officer made a formal decision only about records held on the computer. The outsider status of the researcher was problematic. The case of deceased patients (which is not covered by the Data Protection Act) would have to be decided by the research ethics committee. | No formal decision at 7 weeks |

Predictably, it emerged that patient consent would be required (albeit that this view was not held unanimously). However, it was already becoming clear that time restrictions would render the collection of data from medical records unfeasible, not least because the pilot phase had also demonstrated that the actual data collection process by myself would take an impossible amount of time (and in some centres, data might not even be available).

Hence, the final question to data protection officers resulting from the final protocol change was whether the use of ERCF data would be permissible in

anonymised form, whereby the de-coding key would be held in the centres. This would allow de-coding and checking of the data by internal staff in the centres (see section 4.3.2 for detailed procedures). I addressed the same question to the MREC and to the UK Data Protection Commissioner's Office and their responses were made available to data protection officers. The ultimate decision was thus left with the participating NHS Trusts.

The response letter from the Data Protection Commissioner's Office (Appendix A) emphasised her view on the nature of "pseudonymised" data, which she considered to be personal data and thus under the remit of the Data Protection Act 1998. Her reason for this was the existence of the practical means for identification of individuals, albeit that these means are not in the possession of the recipient of the data (in this case myself as the researcher).

Finally, four out of the five hospital Trusts concerned (covering five of the seven centres) decided that the study was permissible in the proposed form without further consent, given their reading of the data protection legislation and advice. One of these Trusts saw it necessary to request the University who employed the research team to indemnify the Trust in respect of any material risks in the event that the decision was ever challenged - a rare step at the time but now almost standard.

The fifth Trust, after internal deliberations between the Data Protection Officer, Caldicott Guardian, R&D Director, and the Trust's authorised data protection signatory, decided that the retention of a code was not acceptable without patients' consent. The Data Protection Commissioner's letter was interpreted as requiring specific information to data subjects on the particular research so that explicit consent could be obtained. Consequently the local clinician had to destroy the de-coding key he had received from the ERCF operating company, and data quality checks could not be performed in that

Trust. The use of the thus fully anonymous data from the database was seen as permissible.

Thus, four Trusts retained the decoding key in order to identify their patients for data checking. The staff to be involved in decoding the data were selected by the responsible clinicians on the basis that they would have access to the data in the course of their normal duties. Those members of staff signed a form committing them to observing patient confidentiality. I signed a declaration committing me to holding the data securely and not seeking access to the decoding key. On completion of the project, I specifically requested the involved Trust staff to destroy the decoding keys.

### 3.3.4 Concluding Remarks

The difficulties described above were to some extent time-dependant; further relevant guidance documents have been published since, but the debates on the implications of the Data Protection Act, the Human Rights Act, and the Common Law Duty of Confidentiality on medical research have continued in the medical literature (and the publication stemming from this work has frequently been referenced). Despite it often being said that the matter is ultimately for the courts to decide, there has been no further relevant case law.

The Dornase Alfa Case Study itself was severely affected by the uncertainties around the interpretation of the - at times conflicting - legislation and available guidance. Suitable consent arrangements are unlikely to have been put into place at the initiation of many existing databases and registries, and researchers may therefore continue to risk inconsistent interpretation of the guidance by different participating organisations. Evidently, even researchers setting up new registries now are still overwhelmed by the complexity of the

relevant legislation (Haynes *et al.* 2007). However, this seems now inevitable, unless the legal framework is altered to enable important research to take place more easily.

Approval for the use of identifiable data without consent can be sought from the National Information Governance Board (NIGB) for England and Wales (formerly Patient Information Advisory Group, PIAG). However, this mechanism is itself complex and by its very nature intended to be temporary, as approval is conditional on efforts being made to obtain consent in future. Disease registries such as the ERCF need to adapt their procedures to obtain explicit consent from current and future patients in the long run. However, this does little for making already existing databases available for research.

The issue of access to existing data is of significant importance for a vast range of health services research as well as health technology assessments. Patients and the public may still be largely unaware of the implications of the current laws, and it has to be asked whether the current restrictive position serves the public interest. The Discussion chapter sets this issue into the current context.

# 4 Dornase Alfa Case Study: Further Methodological Issues

## 4.1 Introduction

A main concern in ESDs centres around the quality of the available data. So it was in this study. This chapter describes the work undertaken to assess the quality of the database and its data. It explores the dataset and variables within it, and reports the conduct and results of a detailed external verification of key variables. This was later published as the first such data verification of this data source (Strobl et al. 2003), and I was later asked to present this work at a subsequent international CF conference as an invited speaker. This underlines the great interest which the CF research and clinical community awarded to this work. In the last section in this chapter I present an assessment of the representativeness of the available sample from the database. This is in recognition of the oft-made claim that results from observational studies are more generalisable than trial results.

## *4.2 Data Source and Data*

### 4.2.1 The Epidemiologic Registry of Cystic Fibrosis (ERCF)

The Epidemiologic Registry of CF (ERCF) has been set up in 1994 by F. Hoffmann-La Roche across several European countries as a multi-centre, longitudinal, follow-up project of CF patients receiving routine care. The intention of the ERCF was to analyse progression of lung disease by pulmonary function and infection rates (Anonymous 1994). The database monitored lung function, respiratory infections, potential risk factors for lung function decline, and the safety of dornase alfa treatment.

The ERCF had over 13,600 patients enrolled and in 1999 reported a mean observational period of 2.3 years. Centres contributing data on their patients were paid for doing so and might have seen this as a welcome generation of a little extra income. An Advisory Committee including clinicians from participating centres in different countries controlled the ERCF. The management of the Registry was entrusted to a Swiss company, NSPM (Lucerne), whereas the data were held in Frankfurt by Quintiles, a pharmaceutical services company. It received anonymised data on patients from participating CF centres by remote data entry (in the early years of its existence, it received data via paper-based report forms). Only the units themselves could link codes used by the ERCF to individually identifiable patients. All patients of a participating centre were meant to be registered, both those *on* and *off* dornase alfa.

The ERCF Advisory Committee controlled access to the data. Whereas previously no outside research team had sought access to the dataset, the Committee was keen to enable such access. The Committee had hoped to

secure public funding to continue the ERCF after the financial support from Hoffmann-La Roche had ceased. Unfortunately no further funding was granted and the ERCF stopped collecting data in the year 2000.

Several journal papers, regular annual reports, as well as a number of conference abstracts had been published based on ERCF data. Some of the papers presented descriptive analyses (Delaisi *et al.* 1998), but many reported exploratory analyses on different topics: e.g. factors associated with lung function decline (Mastella *et al.* 1999; Navarro *et al.* 2001), practice patterns by age and severity of lung disease (Koch *et al.* 1997), or disease manifestations by different classes of mutations (Koch *et al.* 2001). Conference abstracts included exploratory analyses on the impact of dornase alfa on the frequency of exacerbations and $FEV_1$ decline (Hodson *et al.* 1998b; Hodson *et al.* 1999), and the impact of diabetes on lung function and nutritional status (Koch *et al.* 2000), and descriptive reports on centre and country differences (Harms *et al.* 1998b), and bacterial colonisation rates (Mastella *et al.* 2000).

#### *4.2.1.1 ERCF objectives*

The explicit objectives of the ERCF were as follows (Anonymous 1994):

- *"To collect information on the safety of long-term treatment with dornase alfa and to examine trends in pulmonary function and rates of pulmonary exacerbations that relate to the effectiveness of long-term treatment with dornase alfa and in the patient population who are not receiving dornase alfa;*
- *To further define the clinical course and natural history of specific populations of all subgroups of CF patients using information on concurrent medical conditions, non-routine hospitalisations, deaths, and other data;*

- *To describe practice patterns in the treatment of pulmonary disease in different populations of CF patients including non-routine hospitalisations, clinic visits, routine medications, administration of intravenous, inhaled, and oral antibiotics, and dornase alfa treatment information, if applicable."*

## 4.2.2 Access to the dataset

One of the ERCF Advisory Committee's concerns was the protection of the identity of individual centres/clinicians. Thus, it was stipulated that the lead clinicians in participating CF centres had to give written permission for their data to be made available for the current research study. In order to facilitate the data quality review, lead clinicians in the seven participating centres permitted the research team to receive the code which identified their respective centres. Even though there may not be any legal reason for the centres to remain anonymous, the potential sensitivities arising from comparing clinical data of different centres warranted this cautious approach.

A second layer of anonymisation was added to the already coded patient data for security reasons. Thus each participating centre clinician received a new decoding key from Quintiles, in addition to the one already held routinely.

## 4.2.3 Data management

### *4.2.3.1 Outline of the dataset*

The dataset received from Quintiles arrived as a set of 13 password-protected files in SAS format. Each of the 13 files related to a particular subject (e.g. diagnostic, demographic, clinical, medication, or transfer data), and contained a varying number of reports for individual patients. These were identified by a patient ID number, an enrolment or visit/reporting date, and

numbers distinguishing the reporting and current centres. The dataset spanned the observation period from the beginning of the database in 1994 to the end of the year 1999. Participating centres had been requested to contribute data on all their patients to the ERCF.

#### *4.2.3.2 Data handling and documentation*

Copies of the original files were converted into SPSS format; data cleaning and data quality review were undertaken using SPSS Version 9.0. For some auxiliary steps and for generating feedback forms for the data quality review, Microsoft Excel 2000 was used.

Log files were produced to document the data cleaning of each individual file; later, a set of summary log files was produced and finally a set of variable summaries for each file, detailing for each available variable the data origin, coding, data quality and any alterations or preliminary processing such as re-coding.

#### *4.2.3.3 Security*

Data were stored on the original compact disc received from Quintiles and copies of ongoing working files at different stages were kept. Copies of this material were also saved regularly. On completing the data cleaning and data quality review (see subsequent section), a final full version of the dataset was saved, before reducing and merging files ready for the analysis.

The electronic transmission of files containing BPID numbers and centre codes between research staff was not permitted.

### 4.2.4 Variables

The following list of main variables corresponds to 12 of the 13 files (the remaining file contained data on prompts for further responses from the main recording form):

**Table 4.1: List of data files and variables available in the ERCF**

| Files | Variables |
|---|---|
| Demographics: | Month and year of birth<br>Sex<br>Race |
| Diagnostics: | Year of CF diagnosis<br>Clinical indications of diagnosis<br>Sweat test<br>Genotype |
| Dornase alfa: | Start and discontinuation dates<br>Dose per day<br>Timing of use in relation to physiotherapy<br>Nebuliser/ compressor<br>Doses missed in past 7 days |
| Lung function and spirometry (actual measured): | $FVC$, $FEV_1$, $FEF_{25-75\%}$, RV, TLC (including test dates)<br>(including baseline measurements prior to dornase alfa initiation) |
| Antibiotic use: | Drug name<br>Route of administration<br>Indication (exacerbation or prophylaxis)<br>Start and stop dates |
| Microbiology: | Organisms present in respiratory cultures (with culture dates) |
| Medical history: | Certain conditions / illnesses occurring in the past and during the observational period: e.g. asthma, haemoptysis, ABPA, MIE, gall bladder disease, diabetes, nasal polyps, heart failure, organ transplants, etc. |
| Clinical status: | Purpose of visit<br>Height<br>Weight<br>Cough frequency<br>Sputum productivity, volume, and colour<br>Physical findings (crepitations, wheezing, clubbing, hyperinflation)<br>Blood gases |
| Routine therapies: | Scheduled medications and therapies: e.g. airway clearance, bronchodilators, corticosteroids, diuretics, insulin, mucolytics, NSAIDs, oral supplements, tube feeding, pancreatic enzymes, vitamins, etc. |
| Adverse events: | Non-routine hospitalisations (causes, admission and discharge dates)<br>Deaths (causes and dates) |
| Transfers and discontinuations: | Dates of transfers and new centre number<br>Discontinuation dates and reasons |
| Blood tests: | Dates and results for WBC (white blood count) and IgG tests. |

## 4.2.5 Definitions of key variables and their categorisation

### *4.2.5.1 Date of birth and age-related variables*

Only the month and year of each patient's date of birth were available for security reasons. In the calculation of all age-related variables therefore the day was set to the 15$^{th}$ of the month (and year) of birth. Age-related variables have been categorised into 5-year age bands and into age bands comparable with those of other published studies. Unless stated otherwise, completed life years have been used as cut-off values for categories (e.g. category 5<10 encompasses all cases between age 5.0 and 9.9, rather than 9.49).

### *4.2.5.2 Genotype*

The categorisation of genotype has been undertaken in two ways: (1) into severe ("A"), and less severe ("B") categories, including a category "U" for unknown genotypes, and (2) reporting specifically the occurrence of DF508, other, and unknown or unreported mutations. A geneticist on the project steering group undertook the first A-B-U categorisation. However, only some 3% of the sample were categorised as "B", making a comparative analysis difficult.

### *4.2.5.3 CF Diagnosis*

Everyone included in the ERCF was assumed to have CF. Data from four patients explicitly identified on the database as not suffering from the condition were removed from the dataset.

Only the year of diagnosis was collected by the ERCF, rather than the exact date. Hence, to calculate age at diagnosis the "1 July" of the year of diagnosis was used. Therefore, age at diagnosis has to be interpreted with a half-year margin of uncertainty.

#### *4.2.5.4 Predicted lung function*

Lung function (forced expiratory volume in one second ($FEV_1$), FVC, and other less frequently reported measures) was recorded on the ERCF as actual values in litres. However, since lung function is influenced by a number of factors such as age, gender, height, prediction formulae have been developed which use a combination of these factors to produce reference values of e.g. $FEV_1$ for a patient with a given set of values for the relevant influencing factors. Thus an individual's predicted value can be used together with the actual value to calculate the "per cent of predicted" value.

Two sets of prediction formulae were used in this study. The UK Cystic Fibrosis Foundation research centre based in Dundee uses different formulae for patients aged less than 18 years, and those aged 18+ years (with different sets for males and females). The US Cystic Fibrosis Foundation uses a different formula with separate child, adolescent and adult formulae for each sex (Knudson *et al.* 1983). These two types are referred to as "Dundee" and "Knudson" respectively (for details of formulae used see Appendix A). Unless stated otherwise, the predicted values are categorised into <40% of predicted, 40-<70%, and 70% and over.

#### *4.2.5.5 Height, weight, body mass index (BMI)*

The ERCF recorded height in centimetres. If two reports were available from the same day, the mean value was used. In patients aged over 18 years, missing height readings were replaced by the most recently recorded reading, as long as the patient was over 18 years old at the time of that reading (referred to later as "imputed" height values). Weight was reported in kilograms.

Height and weight were summarised by calendar quarter. Body mass index was calculated on the conventional basis (weight in kilos divided by the square of the height in metres) on these averaged figures.

Anthropometric data were standardised to the UK Child Growth Foundation 1990 references.

#### *4.2.5.6 Dornase alfa use*

Reported start and stop dates of dornase alfa use were used to categorise patients into "never", "intermittent" and "continuous" users. Intermittent users are those who have used dornase alfa intermittently since their first initiation or have since stopped, continuous users are reported to have received the treatment without interruption since their first indication of use. In addition, intermittent and continuous users were distinguished in the analysis depending on whether they started using dornase alfa before or after enrolment to the ERCF.

#### *4.2.5.7 Exacerbations*

A "pulmonary exacerbation" was defined for the purposes of the ERCF as *"a respiratory tract infection that is treated with intravenous, inhaled, or specific oral antibiotics and which excludes prophylactic use."* (Anonymous 1994).

However, exacerbation reports appeared in three different files (forms) in the ERCF (see section 4.3.4.3). All explicit indications of exacerbations were accepted as such and data from the three source files were merged by the start date of a hospital admission for exacerbation, or antibiotic treatment episode for exacerbation, or a clinic visit date for an exacerbation. The last available stop date (either end of course of medication or discharge date) for each start date was accepted as the end of an episode. On the advice of clinicians on the steering group, episodes thus generated which were

separated by less than one week were treated as a continuous exacerbation episode.

### 4.2.6 Concluding Remarks

This brief section has described the data source, the ERCF, and the process for accessing the data, as well as the key variables relevant to the Dornase Alfa Case Study. There was an important political dimension to accessing the data. Whereas there was a clear intention to release data to outside researchers for analysis (not least in order to secure the future of the registry), the ERCF organisers did not have any existing mechanisms for doing so. This is understandable, given that the pharmaceutical company marketing the study drug was the original funder of the registry. It is likely that other industry-funded post-marketing surveillance registries would not be as willing to subject their data to outside researchers pursuing their own agendas - and to checking their quality as rigorously as I did as part of this work (which is the subject of the next section).

The registry collected several dozen variables. The key variables have been introduced here, and the following section is devoted to a detailed assessment of the quality of these variables.

## 4.3 Data quality assurance processes

The ERCF itself had in place reputedly intensive validation processes which involved automated data checking of certain variables and requesting corrections from centres. However, on hearing the reservations of the local Steering Group about data quality, I decided to undertake detailed data cleaning and data quality review processes myself.

The main intention of the Dornase Alfa Case Study project was to use the available dataset for an economic evaluation; therefore, the data quality needed to be assessed and improved by correcting and amending data where necessary. Specifically, the completeness and accuracy of the data, and the comparability of cases across centres needed to be examined.

I used three processes to assess and improve the dataset: first, thorough data cleaning processes, secondly, a detailed data quality review (DQR) to verify the main variables against existing data sources in CF centres, and thirdly, I undertook semi-structured telephone interviews with clinical and clerical staff in the centres to identify relevant clinical and recording practices that may impinge on the comparability of records from different centres.

### 4.3.1 Data cleaning

I formulated the following steps for the data cleaning process. Some of these steps involved simple descriptions of the available data on each variable, others aimed at validating the data through cross-checks between variables, and others still involved more direct judgement by myself of the likely risk of inconsistently applied definitions for variables and hence bias or error in the data:

1. Identification of variables in the data files vis-à-vis the original data entry forms
2. Assessment of:
    a. Completeness of variables
    b. Range and spread of values
    c. Frequency of occurrences / numbers of readings
3. Judgment on potential reliability issues
4. Identification and deletion of duplicate reports
    a. Genuine duplicates (which could be deleted without any loss of information)
    b. Problematical duplicates (where discrepancies between reports exist)
5. Validity of values

    Assessed against other variables in the same or different files of the dataset (e.g. age vs. height)
6. Coding
    a. Previously un-coded text variables
    b. Merging of different variables for easier data handling
7. Identification of questions to be addressed at local centres regarding their practice of recording the data as well as their normal clinical practice.

#### *4.3.1.1 Identification of variables*

Quintiles provided the following documentation:

(1) "Code lists" for all numerically coded variables;

(2) "Schemata" for each file, which contained the variable name, data type, code list, a very brief description (i.e. full name of the variable), the output format, database format, and an indication

of whether the variable was derived and whether it was mandatory;

(3) Copies of all data entry forms denoted with the relevant variable names as well as the form-IDs found in the relevant files (see Appendix A).

Together with a copy of the folder containing all the instructions supplied to local staff entering data into the ERCF, this information enabled me to recognise the variables in the 13 files, link them to the original questions on the data entry forms, and to review the instructions given to staff at participating centres.

This initial assessment was complicated, as two different versions existed of most data entry forms, due to an earlier revision. For some variables, the questions on the forms had not changed. In some cases, data from both versions of the form had been entered into the same variable. In many cases, however, variables contained data from either one or the other form version.

#### *4.3.1.2 Further assessments*

**Completeness of variables and frequency of occurrences**

In each file, an overview had to be gained over the basic numbers of available reports and their completeness: broadly, this included an assessment in each file of the number of reports per individual, the number of reports from each centre, and the number of reports which were complete for each variable.

**Range and spread of values**

For each variable, the range of values and their frequency and spread was assessed. Impossible values had to be deleted and recorded, set as "missing values" so they would be excluded from analysis. Values were also assessed against values from other variables (see below).

#### *4.3.1.3 Judgment on potential reliability issues*

Issues which may have implications for data reliability include:

- ❖ Differences in question wording and question layout between the different forms, where data were coded into the same variable;
- ❖ Variables which were difficult to assess or allowed a significant degree of subjectivity in their interpretation (e.g. daily sputum volume);
- ❖ Variables for which no or no sufficiently unambiguous instructions existed for data entry staff;
- ❖ Variables which originated from a "tick-box answer" (e.g. "check if none").

#### *4.3.1.4 Identification and deletion of duplicate reports*

**Genuine duplicates**

Most files contained duplicate reports, sometimes multiple reports, whereby BPID number, the date of the report and all variables (apart from administrative variables) were duplicated. These reports can be identified easily by using the "AGGREGATE" function in SPSS using the key variables. In most cases, duplicates were deleted manually; only for two very large files (reporting lung function and antibiotic use) with a great number of duplicates (often due to missing data), a "MACRO" was written and applied in Excel for this purpose.

**Problematical duplicates (where discrepancies between reports exist)**

Having used the "AGGREGATE" function for identifying duplicates and assessing them manually, it emerged that not all duplicates were full clerical duplicates. In cases where there were discrepancies, rules for alterations and deletions of those duplicates had to be defined. Generally, where there was no other means of checking the true value of a numeric variable, data was considered to be missing. Discrepant entries in duplicate reports of a text

variable were usually "complementary" rather than contradictory and could therefore be merged.

#### *4.3.1.5 Validity of values*

Wherever possible, values of individual variables were cross-tabulated or cross-referenced with those from other variables. Initially this cross-referencing was restricted to variables from the same file, e.g. variables representing subgroups of a particular variable, or reporting date vis-à-vis event date. In a later stage, cross-referencing was undertaken between variables of different files, allowing e.g. the comparison of clinical data such as height, weight, lung function with age and sex, the use of antibiotics in hospital vs. reports of hospital admissions, or the prescription of insulin or hypoglycaemics vs. the diagnosis of diabetes, or simply the matching of key dates (such as date of birth, diagnosis, enrolment, transfer, death, and discontinuation).

The assessment and alteration of any data followed documented rules. Incompatible values had to be logged and deleted, because in many cases, alterations could not be made with any degree of certainty. However, in cases where more than two variables could be expected to match, this was easier (e.g. if height and age were compatible with each other, but not with the given weight, or if a reporting date, admission and discharge date were all available and the admission date preceded the reporting date by exactly one year). Equally, the timely sequence of values for an individual could be used to make decisions on the likely correctness of individual entries.

#### *4.3.1.6 Coding*

**Previously un-coded text variables**

In many cases, text variables were already coded. For some free-text responses, coding still had to be undertaken. Coding schemes of clinical data

were checked with clinicians and amended following their advice. Entries relating to "other, please specify"-response options were coded also.

**Merging of different variables for easier data handling**
Some variables could be summarised into a smaller number of variables by re-coding them.

#### *4.3.1.7 Identification of questions to be addressed at local centres*

It became obvious that the interpretation of some data would depend on local recording conventions, as well as the clinical practice and case management in each centre. Relevant questions clarifying these were gathered during the data cleaning process and were raised with local clinicians and data entry personnel.

### 4.3.2 Data Quality Review (DQR)

Over and above the data cleaning process described in the previous section, a detailed data quality review (DQR) against original patient records in the centres was possible for five of the seven Northern UK centres (one Trust covering the remaining two centres did not permit the patient identification necessary for the DQR - see previous Chapter). Such data verification exercises are not the rule for database studies, as they involve considerable time and resource, but we saw it as important here, given the doubts expressed by Steering Group members about the quality and completeness of the ERCF data. I was able to publish this work as the first external verification of this dataset (Strobl et al. 2003), and was later invited to present it at in international CF conference (along with findings from the comprehensive Review reported in Chapter 6), underlining the contribution this work made to the scientific and clinical CF community.

#### *4.3.2.1 Selection of variables*

Because of the potential workload involved, the DQR had to be limited to a priority set of variables: some key variables were checked for the total patient population ("population-DQR"), and a further set of variables (demographic variables, genotype, use of dornase alfa, $FEV_1$ and FVC readings, hospitalisation) was checked for a sample of patients only ("sample-DQR"). The choice of variables for inclusion in the DQR derived from their inclusion in the objectives of the study, and also by judgments on the accessibility of data locally for comparison purposes.

**Table 4.2: Variables included in data quality review**

**Key variables checked in the population-DQR:**

- Enrolment centre
- Enrolment date (*for orientation only*)
- Current centre
- Discontinuation date
- Death date
- Cause of death
- Transfer date
- New centre after transfer
- Last reporting date (*only to be checked if <1999*)

**Further variables checked in the sample-DQR:**

- Month and year of birth
- Sex
- Year of diagnosis
- Genotype
- Enrolment date
- Discontinuation date
- Reason for discontinuation
- FEV1 and test date
- FVC and test date
- Indication of whether spirometry was pre-dornase alfa baseline value
- Dornase alfa therapy (yes/no)
- Dornase alfa therapy start and stop dates
- Adverse events
- Hospitalisations
- Admission and discharge dates

#### *4.3.2.2 Review process and pilot*

The five participating centres had been supplied with decoding lists by Quintiles, allowing them to identify their own individual patients. Thus a member of each centre's staff could be supplied with anonymised data by the research team and check their correctness against other existing in-house records and data sources.

Each reviewer received Summary Lists for the ***population-DQR***. These Summary Lists were split into

1. Current patients;
2. Patients enrolled at the centre but transferred while on the ERCF.

The purpose of those lists was to assess the completeness of the database, with a view to transfer, discontinuation, and death data (see Appendix A for a sample of data feedback forms).

The ***sample-DQR*** comprised 4 forms for each patient record, one for once-only recorded items (demographic, diagnostic, and follow-up details), and one form each for spirometry readings, dornase alfa treatment details, and hospitalisations.

In addition to written guidance on how to check the data, I visited most reviewers on site to explain the process.

For a small ***pilot*** of the sample-DQR, six records were checked in each centre. These were chosen to represent a spread of different lengths of observation: one record of a patient "on" and one of a patient "off" dornase alfa of a length of follow-up of 1, 3 and 5 years (or the next available record closest to these periods), avoiding transferred patients.

#### *4.3.2.3 Selection of records for sample-DQR*

The time taken for these 6 records was used to calculate an average length of time needed to review one record in a centre. Because of time and funding restrictions, a pragmatic decision was made to determine the size of the sample to be reviewed in each centre depending on the amount of time each person was able to commit to the process within the available 10 weeks.

The record samples were made up of (a) records for which discrepancies or queries regarding the key variables had been identified in the original data cleaning process, and, if this number did not already exceed the estimated possible number of records a local person could review, (b) a further number of randomly chosen records from each centre to make up the estimated manageable number.

#### *4.3.2.4 Amending data files*

Additions, deletions, and any amendments of the existing data files were undertaken manually according to the data reported back by the local DQR-reviewers. Their corrections were given preference over existing data. Each amendment was logged on a separate Excel spreadsheet. A separate Excel sheet was created later to collect detailed information on the nature of any necessary amendment or correction; at this stage, the first spreadsheet was double-checked.

### 4.3.3 Semi-structured interviews

The review of the dataset and data cleaning process identified a number of questions regarding recording and clinical practices in the different centres. These questions were collected and used to formulate a semi-structured schedule of questions, which were then addressed to clinicians and centre staff responsible for recording ERCF data via telephone.

## 4.3.4 Results of data quality assurance processes

This section reports the results of the data quality assurance processes and discusses their implications for the Dornase Alfa Case Study.

### *4.3.4.1 Completeness of data*

**Completeness of reports**

Differences in reporting rates for key variables might be used as a crude indicator of problems with data completeness. In this regard, the proportion of each centre's patients for whom reports existed in data files referring to clinical status, microbiology, and antibiotic use was similar across all centres (97-100%) (see Table 4.3). In contrast, the proportion of patients for whom entries existed in the file relating to non-routine hospitalisations/adverse events (25-74%) varied, as did the proportions in the file on blood tests (0-97%). This variation suggests that recording of these variables is unreliable; blood tests did not seem to be reported at all in some centres.

**Table 4.3: Proportion (%) of patients per reporting centre appearing in each variable group (after data cleaning)**

| Variable group (file) | Centre and number of records associated with each centre | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Centre 1 | Centre 2 | Centre 3 | Centre 4 | Centre 5 | Centre 6 | Centre 7 | Total |
| | **161** | **137** | **200** | **181** | **119** | **286** | **215** | **1299** |
| Clinical status | 100 | 100 | 100 | 100 | 100 | 99 | 100 | 100 |
| Medical history | 100 | 100 | 100 | 100 | 99 | 99 | 100 | 100 |
| Routine therapies | 97 | 100 | 99 | 100 | 99 | 99 | 100 | 99 |
| Dornase alfa use | 100 | 100 | 100 | 100 | 99 | 99 | 100 | 100 |
| Pulmonary function | 82 | 55 | 77 | 81 | 99 | 98 | 99 | 86 |
| Micro-biology | 100 | 99 | 97 | 100 | 98 | 98 | 98 | 98 |
| Blood tests | 84 | 2 | 18 | 84 | 97 | 55 | 0 | 46 |
| Adverse events | 57 | 19 | 25 | 47 | 70 | 74 | 60 | 52 |
| Antibiotics | 99 | 100 | 100 | 100 | 97 | 98 | 97 | 99 |

Notes:
a) Individual patients may appear in more than one centre.
b) Centres 1-4 are children's centres, centres 5-7 are adult centres.
c) Centre 2 only contributed to the ERCF for two-three years (other centres: six years).
d) Centres 4 and 7 did not take part in the DQR.

Interviews with centre staff highlighted that the frequency of routine clinic appointments varied between and within centres. Most centres saw patients every 2-3 months, but some operated a more flexible review system by inviting patients depending on their individual need and circumstances. Thus, the frequency and regularity of reports would have varied. Many centres conducted more or less standardised annual reviews of all patients, and most reported that there would be little if any variations in practices between doctors of the same centre.

However, it was also elicited that the number of visits recorded on the ERCF does not represent the actual number of visits. Local ERCF coordinators reported that they tended to adhere to an upper limit of four visits per patient per year which were entered into the ERCF. Some selected four visits

arbitrarily, others tried to space them reasonably equally throughout a particular year. Overall, the number of reports for a patient did not correspond to the number of quarter-years under observation.

This indicates considerable differences in reporting frequency and practice between centres and considerable potential for bias through under-reporting certain key variables such as adverse events, i.e. a key outcome! This means that the dataset should not be analysed as one homogenous sample, but centre differences need to be taken into account in the analysis, at least of some if not all variables. More worryingly, there are reporting practices at play which the ERCF operators would most likely not have realised, but which would have a significant impact on the eventual quality of data and reliability of any subsequent analysis and interpretation.

**Completeness of patient records**

Some centres have provided overall numbers of their patient populations and information on how many of those patients are registered on the database. This information could be compared with the numbers of individuals found registered on the dataset. In three centres the numbers matched closely at the end of any particular calendar year (up to 8 patients missing in the ERCF dataset); different date points being chosen for the count may explain some of that variation. However, underreporting definitely affected one centre, where registration stopped after a certain number of patients were reached. In one other centre the ERCF dataset seems to contain more patients than were reported verbally for a particular year, probably due to shared care patients being included in the register. For two centres no comparison data existed, but local ERCF coordinators were confident in a total coverage of their patient population, except for 1999 when registration and reporting ceased.

Of one of the centres not included in the DQR, summary data were available, which could be compared to the data held in the ERCF. This was particularly interesting, because this centre reported that only part of its patients were registered on the ERCF. Summary data and comparative values from the ERCF sample are presented in Table 4.4. Whereas some of the clinical data are well comparable between the two sources, the prevalence of bacterial colonisation, but also the use of dornase alfa seem under-reported in the ERCF data. It is unclear to what extent these differentials might be explained by non-registration of more severely affected patients, or incomplete reporting on recorded patients. However, the relevant clinician reported that the ERCF data relating to the centre was less reliable than data held in the centre's own CF database (the source of the summary comparison data).

**Table 4.4: Available summary data from one centre compared to data of the same centre held in the ERCF sample**

| | Summary data available elsewhere | Data available in ERCF sample |
|---|---|---|
| Mean $FEV_1$% of predicted between 1995 and 1999 | 55 | 50 |
| Mean body mass index (1999) | 21.2 | 20.9 |
| Prevalence of diabetes (%) (1999) | 28 | 27[1)] |
| Dornase alfa use (%) (1999) | approx. 40 | 31[2)] |
| Number of patients in 1999 | 262 | 196[3)] |
| % of patients with $FEV_1$<40% of predicted (1999) | 27 | 32[4)] |
| % of patients reporting Pseudomonas aeruginosa (1999) | 76 | 50 |
| % of patients reporting Burkholderia cepacia (1999) | 9 | 6 |
| Mean age (years) (1999) | 25 | 25.4[5)] |

[1)] Includes all patients with any diabetic therapy recorded.
[2)] Cannot include patients who continue therapy during 1999 but have no report of use in 1999.
[3)] 2 patients from that centre were amongst those excluded because of missing demographic data
[4)] Any reading of <40% predicted $FEV_1$ is included, not mean during entire year 1999.
[5)] Mean age (on 31.12.99) of patients on register at end of 1999.

### *4.3.4.2 Records subjected to sample-DQR*

A total of 230 case records were subjected to the Sample-DQR process in five centres; the number of records verified in each centre varied (Table 4.5).

**Table 4.5: Numbers of patients in sample-DQR**

| Centre | Total record number (last centre) | Sample for DQR | | |
|---|---|---|---|---|
| | | Records with queries for checking | Additional randomly selected records | Total number of records in sample, including pilot[1)] |
| 1 | 127 | 21 | 14 | 41 |
| 2 | 115 | 19 | 0 | [2)]6 |
| 3 | 179 | 35 | 0 | 40 |
| 5 | 130 | 27 | 13 | 46 |
| 6 | 282 | 81 | 20 | 106 |
| Grand Total | 833 | 183 | 47 | [3)]239 |

[1)] This is not necessarily the sum of the previous two columns and the 6 pilot records, because the latter may have already included case records which contained queries for checking.
[2)] No further case records accepted for checking due to time pressure.
[3)] Nine actual patient records appear twice, e.g. if they were checked in different centres. Hence, a total of 230 case records have been subjected to DQR.

Not all records delivered to the units for checking could actually be reviewed; reasons were as follows:

(a) BPIDs were not recognisable from the de-coding lists the centres had received, because the lists did not contain all the necessary number pairs (e.g. for very recent database entries);

(b) Only part of a patient's medical records were kept on site, either because the patient had been transferred, or was treated under a shared care arrangement and therefore records were kept in a peripheral hospital;

(c) Patients' records were inaccessible;

(d) Variables were not recorded in the notes (e.g. start and stop dates of dornase alfa in one centre).

Table 4.6 presents the completeness of checks undertaken on the sample of records. Of all 230 patients reviewed, 140 were users of dornase alfa at some point in time, but 24 of these were allocated to the centre which could not review the information on dornase alfa use. Thus, dornase alfa use could not be verified for a total of 64 records (45 in one centre), hospital admissions for 23, and lung function tests for 19 records.

**Table 4.6: Completeness of checks in sample-DQR**

| Extent of data verified* | Number of records | % |
|---|---|---|
| All data items | 127 | 55% |
| 75-99% | 40 | 17% |
| 50-74% | 37 | 16% |
| 1- 49% | 11 | 5% |
| None | 15 | 7% |
| TOTAL | 230 | 100% |

*one item = either a lung function, dornase, or admissions follow-up report (including all relevant variables), or an item of fixed data (e.g. sex, genotype, discontinuation date)

### *4.3.4.3 Data quality of main variables*

The following text summarises the findings on the data quality of key variables only resulting from the data quality assessments, preceded by an overview of data disagreements identified prior and through the DQR:

**Disagreements identified in the data**

Table 4.7 summarises the corrections of some key variables which were possible to be undertaken during data cleaning (i.e. without the DQR).

**Table 4.7: Corrections of key variables possible without DQR**

| Variable | Corrections | Total number of reports |
|---|---|---|
| Month and year of birth | 10 (variable appeared in two files, on multiple reports in one file) | |
| Sex | 1 (variable appeared in two files, on multiple reports in one file) | |
| Year of diagnosis | 1 deletion (impossible value) | |
| $FEV_1$ | Deleted 5 values <0.01 and >6, plus 1 severe outlier<br>13 discrepancies in duplicates replaced by mean reading for that test date | 18,464 lung function reports |
| FVC | Deleted 14 values <0.01 and >8, plus 4 severe outliers<br>12 discrepancies in duplicates replaced by mean reading for that test date | |
| Lung function test date | 5 deleted (<1990, but were definite follow-up reports)<br>68 alterations (usually of YEAR in test date) following comparison with report date and other information in files | |
| Admission dates | 4 corrections | 2,766 admission reports |

Table 4.8 summarises the disagreements identified in the sample-DQR between the ERCF data and original patient records. The greatest area of disagreement was on hospital admissions. Both for this (21 additions, 18 in one centre) and for dornase alfa treatment (6 additions), additional reports were found on examining the patient case notes.

**Table 4.8: Disagreements between ERCF data and DQR**

| Variables | Disagreements between ERCF data and DQR reports | | Number of records verified | % of verified records with disagreements |
|---|---|---|---|---|
| ***Sample DQR*** | | | | |
| Sex | 2 | (1 addition) | *214* | *0.9%* |
| Month and year of birth | 6 | (1 addition) | 214 | 2.8% |
| Year of diagnosis | 4 | (1 addition) | 213 | 1.9% |
| | | | **Number of relevant follow-up reports[1)]** | **% of follow-up reports with disagreements** |
| $FEV_1$ | | 45 | 3602 | 1.2% |
| FVC | | 47 | 3602 | 1.3% |
| Lung function test date | | 48 | 3602 | 1.3% |
| dornase alfa (yes/no) | | 11 | [2)]2481 | 0.4% |
| Dornase alfa start dates | | 15 | [2)]2481 | [2)]0.6% |
| Dornase alfa stop dates | | 13 | [2)]2481 | [2)]0.5% |
| Admission dates | | 34 | 925 | 3.7% |
| ***Population DQR*** | | | | |
| Death | 8 | (all additions) | | |
| Death date | 18 | (17 additions) | | |
| Cause of death | 21 | (all additions) | | |
| Transfers | 12 | (8 additions) | | |

[1)] Total of reports appearing in either ERCF or DQR (including those not occurring in DQR).
[2)] A minority of these reports contain actual start or stop date entries; there were a total of 189 start and 99 stop dates concerned; using these figures as denominators would bring the disagreement rates for start and stop dates to 8% and 13% respectively.

This data demonstrates that relatively little actual data correction could be achieved through routine data cleaning only; the majority of such corrections were of lung function test dates. The disagreement rates from the sample DQR ranged from under 1% to 3.7% for admission dates. This is globally speaking a satisfying result. However, it is likely that the ERCF's own data verification processes would have been focussed on similar key variables. Therefore, the relatively low rates may not be indicative of other variables not checked here or by the ERCF. The under-reporting of death - the easiest, memorable and reliable outcome measure of all - by contrast is a significant and rather disappointing finding. This is not unknown in databases, and without specific additional safeguards to ensure completeness of reporting, under-reporting of outcomes remains a significant risk.

**Demographic information**

Both sex and date of birth were recorded twice in the ERCF, once in the demography file and again on each report in the file on adverse events and hospitalisations (for approximately half the patients on the database). Reports from those two files are unlikely to have been completed on the same day, so if there was agreement between both files, then there is good reason to be confident about the reliability of these variables. Whereas there were some wrong entries on the multiple reports in the adverse events file, the majority of them matched the entries in the demography file (one "female" entry against several "male" entries for the same individual can be corrected with a good degree of confidence).

The DQR identified two disagreements on sex entries and six on month/year of birth (Table 4.8).

Five individuals were not recorded in any of the two files mentioned, but reported on in other files. Only one of the five BPIDs could be identified by the DQR-reviewers in the centres and the demographic information completed. The other four could not be included in the sample.

**Transfers and discontinuations**

Twenty-four patients appeared on the database as registered in two different centres with two different ID numbers. In many cases, this dual registration was due to the fact that patients under "shared care" arrangements between a specialist centre and a local hospital had been transferred to another specialist centre, which perceived the local hospital to have been the main treatment centre; whereas in transfers between specialist centres (who know of each other's participation in the ERCF) the ERCF ID numbers were routinely passed on, this clearly did not always happen in transfers of shared care patients. It would appear that there was overall an insufficient mechanism to ensure that links between patient episodes in different centres are made.

In the original dataset, 142 transfers had been indicated (with 141 transfer dates). The DQR revealed a further 8 transfers during the observation period. Four transfers indicated in the dataset were identified (in the DQR) as not having taken place.

**Deaths**

Similarly to demographic data, patients' deaths had been reported in two separate files of the original dataset: the transfer and discontinuation file (which contained a discontinuation date only and a variable to indicate the reason for discontinuation), and the adverse events and hospitalisations file (which contained the death date and cause of death).

Whereas the first file had reported 91 deaths, only 84 of these were reported in the latter file, where a definitive date and cause of death were recorded. All 91 deaths were confirmed in the DQR. The DQR revealed also an additional 8 deaths having occurred during the observation period.

Only one date of death was corrected in the DQR. Twenty-one causes of death and 17 death dates were added during the verification, but several of

the latter had already been recorded on the ERCF as discontinuation dates. The causes of three deaths remained missing, because these deaths occurred in a centre not taking part in the DQR.

This is now one of the most reliable variables, because (a) it was part of the total population-DQR and (b) it is likely that centre staff know which of their patients have died, even without checking their records. However, this conclusion does not apply to the two centres not participating in the DQR. Deaths were under-reported by 8%.

**Exacerbations**

Despite being a key outcome variable, exacerbations were reported at three different places in the dataset:

1. ***Adverse events and non-routine hospitalisations file:***

   Variables describing an adverse event were used to create a new dichotomous variable which identified all reports involving the term "exacerbation" (but not merely "exacerbation of symptoms");

2. ***Clinical status file:***

   "Exacerbation" was an answer option to indicate the purpose of a clinic visit;

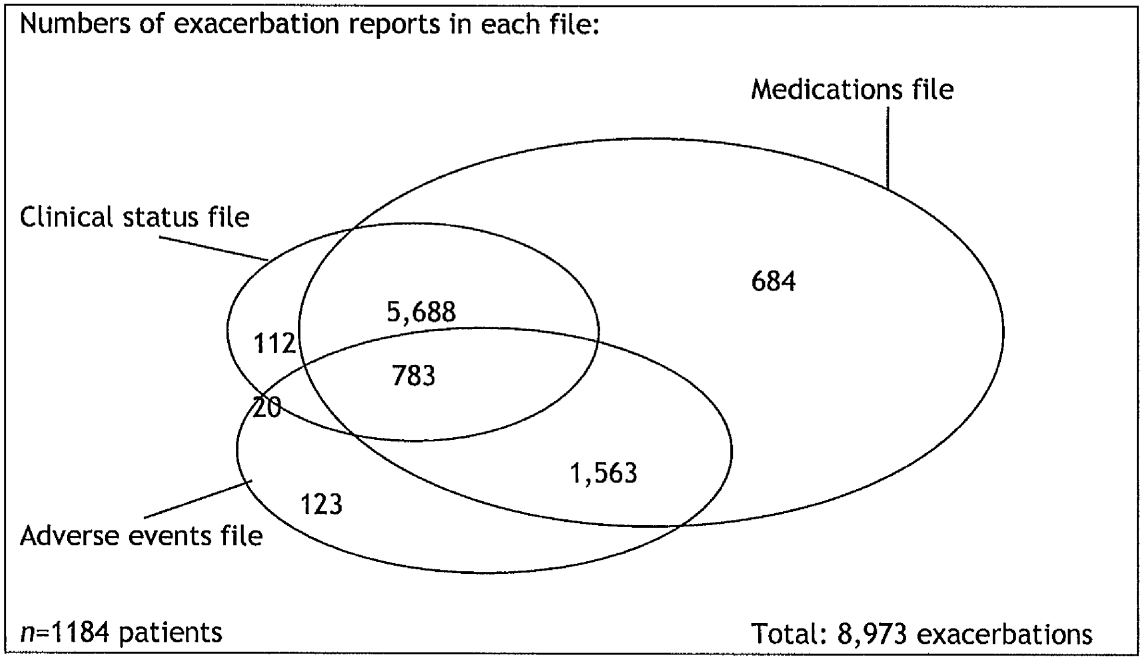
3. ***Medications file (i.e. antibiotic treatments for RTI):***

   "Exacerbation" was an answer option to the question on indication of antibiotic treatment.

The majority of identified exacerbations (97%) were recorded in the medications file. Figure 4.1 gives on overview of the congruence of reporting of exacerbation episodes in the three files. In order to arrive at defined

outcomes, an operational definition had to be agreed with clinical advisors (for the operational definition of exacerbation used see section 4.2.5.7)

**Figure 4.1: Congruence of different sources of data on exacerbations**




The number of exacerbations reported per patient varied considerably between the centres, even leaving aside the centre which contributed to the ERCF for a shorter time. The variation is particularly gross between children's centres, but rates also vary about 2-fold between adult centres. It seems unlikely that all of this variation is accounted for by a real difference in exacerbation frequency, and differences in reporting practices, and particularly under-reporting by some centres seems likely. Note that a large number of reports were lost due to incomplete data on drug names, and treatment start- and stop dates.

No centre seemed to have an agreed centre-specific definition of "exacerbation"; some local coordinators reported that they usually chose the response category "exacerbation" if an increase of respiratory symptoms was observed at the time of a hospital admission or when antibiotics were prescribed over and above the prophylactic regimens and regular treatment courses against *Pseudomonas*.

There were great variations between centres, patients, and also over time in the location of intravenous antibiotics administration (i.e. at home or in hospital). Patients in all centres were routinely on continuous oral antibiotic prophylaxis.

**Hospitalisation**

Similarly, to exacerbations, the database contained three different sources of information on hospitalisation events:

1. ***Adverse events and non-routine hospitalisations file:***

   "Hospitalisation" as a dichotomous variable (specified as "non-routine" in Version II of the form only);

2. ***Clinical status file:***

   a. "Hospitalisation" as an answer option to the question on the occasion of enrolment (specified as "non-routine" in Version II of the form only) on enrolment;
   b. In answer options to the question on the purpose of a clinic visit: "non-routine hospitalisation" on Version II of the follow-up form only, and "planned hospitalisation" on Version II of both follow-up and enrolment forms;

3. ***Medications file (i.e. antibiotic treatments for RTI):***

   "iv- in hospital" as answer option to the question on the route of administration.

In addition, the file on medical history of patients logged the year of an organ transplant; this is known for 24 out of 44 patients having received a transplant; if the year falls into the observation period, the event may also be logged in the adverse events file, but since transplants may well not be seen as non-routine hospitalisation, they may not have been reported there.

In terms of congruence between the data sources, data from the clinical status file (point 2 above), which indicated hospitalisations other than planned hospitalisations, were confirmed by data from the adverse events file (point 1 above): only 10 patients were not covered by adverse events data, 9 of whom are specifically reported to have attended for a "non-routine hospitalisation", and the tenth patient having attended as a "new patient". The agreement between hospitalisations reported in the adverse events file (point 1 above) and intravenous treatment in hospital (point 3 above) was far less satisfactory.

**Dornase alfa therapy**

The database contained a whole host of variables relating to the use of dornase alfa. Given that the daily dose and frequency of use were practically uniform across all users, the main interest from the point of view of this study lay in periods "off" and "on" dornase alfa and characteristics of "never users". Hence, the most important variables were start and stop dates for the treatment and any indications of patients being "off" or "on" or continuing treatment at any given point in time. Therefore, only these variables were included in the sample DQR.

The DQR dealt with 140 patients ever having received dornase alfa, 24 of whom were from the centre unable to review dornase alfa information, and a

further nine records were not or only minimally reviewed. The review reported seven wrong or missed start dates and four such discontinuation dates.

Some patients were on dornase alfa continuously, whereas others simply underwent short trial periods of some two weeks or receive dornase alfa for short treatment episodes only during a hospital admission. For many patients, the sequence of start and stop dates was compatible with supporting variables such as whether a patient was continuing treatment at the time of a report. In a few remaining cases, the sequence was obviously broken, e.g. where there was no stop date between two different start dates for a patient.

A key concern was the compliance of patients with the prescription; whereas one variable (doses missed in past 7 days) clearly aimed to elicit this information, the variable was far from complete. A small study in Manchester and Salford compared cost claims for dornase alfa prescriptions in primary care with expected costs from the start date of the prescriptions for 13 adults and 10 children. Based on this simple comparison, the estimated prescription uptake was 59% for adults and 78% for children (Talbot *et al.* 1998).

All centres reported the use of a protocol to select eligible patients for dornase alfa treatment. In many cases colonisation with *P. aeruginosa* and sputum production and/or cough were amongst the eligibility criteria. Sometimes lung function values (e.g. <70% predicted $FEV_1$) were also considered; patients awaiting transplants may also have been selected for treatment. The duration of individual trial periods was mostly 2 weeks, but longer in some centres or for severe patients (up to 3 months) (Conway 1997; Ledson *et al.* 1998). Some centre protocols included a review after one year of treatment.

Although a date of registration into the study could be identified by the date on the original enrolment form, the patient may well have been on dornase alfa treatment for some time. If a patient was *not* noted as using dornase alfa on enrolment, it was not clear if the patient had been using it previously. There is evidence in the literature that cessation of treatment is followed by rapid diminution of effect, so there may not be any "carry-over" effects. A more serious worry is a selection effect - cessation because of unsuccessful use may indicate that the patient has systematically different characteristics from one who simply has not had a trial of dornase alfa.

Some centres used dornase alfa for short-term symptom relief for in-patients. This typically resulted in (sometimes repeated) short treatment episodes of maybe two weeks. Whereas most centres recorded short-term use and initiation trial periods on the ERCF, two reported not to have recorded these at all. One of them also reported that the start date chosen was at the end of the trial period. These circumstances make it practically impossible for analysts to select pre-treatment lung function values from the database.

**Spirometry: $FEV_1$ and FVC**

About one fifth of individuals with spirometry reports (n=1040) were reviewed. The review reported low rates of discrepancies, and the difference between the original and corrected values was far below 1 litre (actual).

After the DQR, 12 (for FVC) and 13 (for $FEV_1$) pairs of duplicate reports with discrepant readings for the same day had to be merged by using the mean of each pair of readings. Again, the differences were minor.

Predicted values of both variables have been compared with height and age; thus, gross outliers of actual values of spirometry measures could be highlighted and deleted.

However, there were systematic centre-differences in lung function readings. This was confirmed by the findings from interviews: on each data entry form for the ERCF, there was space for two readings of $FEV_1$ and FVC from different test dates, but the selection of readings to be recorded was left up to the centres themselves. Whereas some centres selected any one or two measurements recorded since the previous visit, one centre chose to record the best and worst readings from the period since the previous visit. Another centre recorded the pre-lung function values and yet another centre selected a "good" reading, and thus would have chosen a reading *after* a course of treatment rather than a reading preceding treatment.

**Height**

Height is needed to calculate predicted spirometry measures. It was recorded for all patients on enrolment, and for children under the age of 18 at every visit; for all but nine patients there existed at least one reading.

Height was recorded in two different files: the clinical status file and the lung function and spirometry file; hence, similarly to the demographic information, the two files allowed some collation of the reports. A merged file was created, merging all readings for the same date by their arithmetic mean. This method took care of any discrepancies between the two files, after having first assessed their size (after data cleaning, only one report had a remaining discrepancy of 4 cm in height between two reports of the same day, a further few reports recorded smaller differences). This variable was not part of the DQR.

Informal checks for height against weight and height against age as well as assessments of BMI extreme values (12 or less and 30 or more) helped in identifying outliers, which could then be assessed against the trend of consecutive reports for a particular patient and amended or deleted.

**Genotype**

Genotype for both chromosomes was reported in 74% of patients (892); for a further 10% of patients, the genotype was known of only one chromosome. The DQR was very valuable in increasing the rate of known genotypes and added information on 51 patients (74 chromosomes). Presumably much of the added information was not previously available, or at least not on enrolment. Four entries were corrected during the DQR. The proportion of patients reported to have been genotyped varied between centres (from 73% to 98%).

**Shared care**

The interviews identified that some patients were subject to "shared care" arrangements between a specialist and a peripheral hospital. Such arrangements varied greatly between centres in terms of frequency of contact with specialised CF clinicians, the degree of similarity of disease management practices between patients directly cared for by the centre and "shared care patients", and also their location (i.e. CF centre or peripheral clinic). In some instances, clinical guidelines may be shared with local non-CF clinicians to ensure a consistent approach; in other cases, local non-CF clinicians may consult CF clinicians only sporadically. Only some centres enrolled their shared care patients on the ERCF, others only entered patients for whom they were sole providers of care. This was not a variable in the database, but clinicians were concerned about its possible impact.

## 4.3.5 Discussion and concluding remarks

The ERCF had in place a detailed system of checking the data on key variables and requesting corrections by local centres, if missing or discrepant information was identified. Nevertheless, there are limits to the extent of checking which can be routinely undertaken and therefore further detailed data cleaning and quality checks were undertaken as part of this study.

The Data Quality Review reported here is the only known (and published) such effort concerning the ERCF database. It provides insights into the quality of the key variables meant for inclusion in the analysis and more generally. Many of the problems and issues identified here may also apply to other multi-centre disease databases, particularly those with less intensive routine verification processes than existed for the ERCF!

Even with this limited exercise, important information was gleaned, e.g. eight previously un-reported deaths (8%), much additional genotyping information, and the conclusion that spirometry was relatively well recorded, but outcomes data were badly defined and poorly recorded (deaths, exacerbations, hospitalisations). In addition, the processes for patient transfers missed a considerable number of "shared care" patients.

The DQR generated low disagreement rates on a number of key variables in the ERCF (notably demographic variables, and lung function). Whereas this finding cannot be generalised to the entire ERCF, one may assume that the available subset contained reasonably reliable lung function data. Differences between centres in interpreting lung function reporting rules (i.e. whether a particularly good or poor result was selected for reporting), however, still caused systematic differences between centres.

Chronic disease databases depend heavily on the commitment as well as expertise of local data coordinators and collaborators. Allowing a degree of freedom in defining particular data fields for their own purposes and interests can provide an attractive incentive. Researchers using multi-centre data need to know local reporting rules in order to judge the comparability of data across centres.

Similarly, interviews with centre staff indicated variations in clinical practice between centres which may not be well recorded in a disease registry (e.g. differences in treatment protocols). This risks misattributing variations in outcome to recorded variables when they may arise from unrecorded variables. At the planning of the Dornase Alfa Case Study, clinicians were concerned that outcomes of patients may be partly determined by whether they were looked after under "shared care" arrangements with local non-specialist clinicians. Such "soft" issues may be ignored in large clinical databases.

The issue of motivation to participate in and report to a large clinical database is important. One centre operated its own database alongside the ERCF. It seems thus likely that contributing high quality data to the latter would have been of comparatively lower priority. CF clinicians may have seen the ERCF as a means of generating income for their centres, as they were reimbursed for contributing data. Only if a centre perceived its own purposes to be served by the data collection efforts, might staff be sufficiently encouraged to report complete data and focus on data quality locally.

Incomplete outcomes data pose a problem for particularly economic evaluations, as neither resource use nor health outcomes per se can be estimated reliably unless the extent of under-reporting is known. The ERCF had operated very detailed routine data quality assurance procedures, but under-reporting is difficult to address without reference to other data sources for verification. The DQR found that deaths and hospitalisations were under-reported by some centres (overall by 8% and at least 2% respectively). A verification of data from the US Veterans Affairs' HIV registry also identified underreporting of deaths and difficulties in routinely verifying such data without accessing other data sources (Rabeneck *et al.* 2001).

Whereas only a detailed verification process can identify the extent of under-reporting, comparisons of data from different centres could highlight a range of data quality problems. In the Case Study, centre comparisons could identify problems with under-reporting (e.g. of blood results, exacerbations), and point towards the possibility of reporting differences (e.g. lung function) or important differences in clinical practice (e.g. use of iv antibiotics in hospital).

For a pharmacoeconomic evaluation, compliance with drug treatment is an important issue, both from the point of view of categorising the intervention as well as the associated costs (Hughes *et al.* 2001). A categorical variable which could have estimated compliance (number of doses missed in last week) was poorly completed in the ERCF (over 80% of entries missing). Verification of such a variable would be very difficult, but the issue of measuring compliance is itself complex. Previous studies highlighted that patient self-reporting as a measure of compliance with dornase alfa can be highly unreliable, and compliance has been observed to vary considerably between patients (Dodd *et al.* 1998; Phillips *et al.* 2001b).

Researchers using secondary data have to be familiar with the objectives of the original data collection. For example, post-marketing databases tend to focus on safety and to a lesser degree on efficacy evaluation. In the case of the ERCF, only non-routine hospitalisations and related adverse events were meant to be recorded, however, health economists may have wanted to estimate overall hospitalisation rates for different groups of patients. On the other hand, different centres may be more or less likely to hospitalise patients for the treatment of pulmonary exacerbations, and thus exacerbations per se may be a more useful outcome measure.

However, pulmonary exacerbations were difficult to determine from the data. Fields from several files had to be cross-referenced and assumptions made

about realistic time gaps between distinct episodes. The indiscriminate use of any one field alone (e.g. antibiotic use) would have misrepresented the number of events. It is common to have to use complex algorithms to retrieve data on episodes of care or illness from large databases (Schulman *et al.* 2001). The strength of disease registries is their focus on a single condition, so they ought to enable more precise measurements of key outcome variables than are possible in generic databases. However, for some health outcomes of interest there may not be a clear and unambiguous clinical definition within the medical community (CF exacerbations are a case in point).

The timing of outcome events is important for longitudinal analyses; without timing data, it may be difficult even to distinguish between different episodes. The identified errors in lung function test dates seemed to have been largely due to clerical mistakes, as usually only one item of the day-month-year combination had to be corrected (the year of a date seemed particularly error-prone in dates near the turn of calendar years). Computerised data verification may highlight gaps of one year or more between reporting and test dates for checking.

Some degree of inaccuracy of data in pharmacoepidemiology may be acceptable so long as it is random. Given the low level of inaccuracies found, it was difficult to assess this issue with a sufficient degree of confidence. However, centre variations were an important factor here.

Much of the data cleaning process reported here was based on crosschecks of different variables and essentially a series of basic techniques and meticulous reporting. Such crosschecks allow a greater degree of confidence in the data quality resulting after the cleaning process. In some cases, dual reporting (i.e. the same variable reported in two files) introduced another possibility for crosschecking. For longitudinal records in particular, the time sequence of

readings creates yet another opportunity to aid decisions on the likelihood of individual readings.

Some of the variations in recording and clinical practice between centres elicited here were suspected during data cleaning. Other differences may have gone undetected without the rigorous DQR and questioning process. The examples presented here present potentially significant problems for effectiveness and particularly cost-effectiveness analyses, as variations relate to (recording of) resource use such as clinic visits, and extent or practice of treatment, but also intended effects, as well as exacerbations.

#### *4.3.5.1 Limitations of the DQR*

There are two major problems of data quality in clinical databases: (1) where incorrect entry or lack of entry of available information occurred, or (2) the original source does not reflect the true condition of the patient (Sørensen *et al.* 1996). The focus of the data cleaning and DQR was only on the first of these. Even there, a multitude of causes could have resulted in the observed disagreements between ERCF and original data. Because of data protection regulations, I did not have access to original data and therefore did not have detailed insight into the nature of the data sources used by the DQR reviewers. This limited the exploration of causes for any errors detected and of possible improvements of the data source.

The sample was not a representative random sample of records from all centres, but partly focussed on records which had raised questions during data cleaning. Hence, the proportion of reports with disagreements may over-estimate the true rate of disagreement in our dataset. However, there is no clear relationship between whether records have been found to contain possible discrepancies or missing data before the verification, and their

likelihood of showing disagreements in the verification, probably because only certain types of errors could be identified a priori.

There was no further checking of the validity of the verification itself and it is theoretically possible that some disagreements were missed. In some centres, staff that had been responsible for the original data entry were involved as verification reviewers. In theory, they might have preferred not to highlight discrepancies or errors, but there was no reason to suspect that that was the case.

The selection of variables for the sample verification was limited by the local data sources available for cross-referencing. Thus, it would have been futile to attempt a verification of the variable "exacerbation" as such. The ease of checking varied between local record systems. Future needs for data verification could already be envisaged and planned for at the onset of data collection for clinical databases.

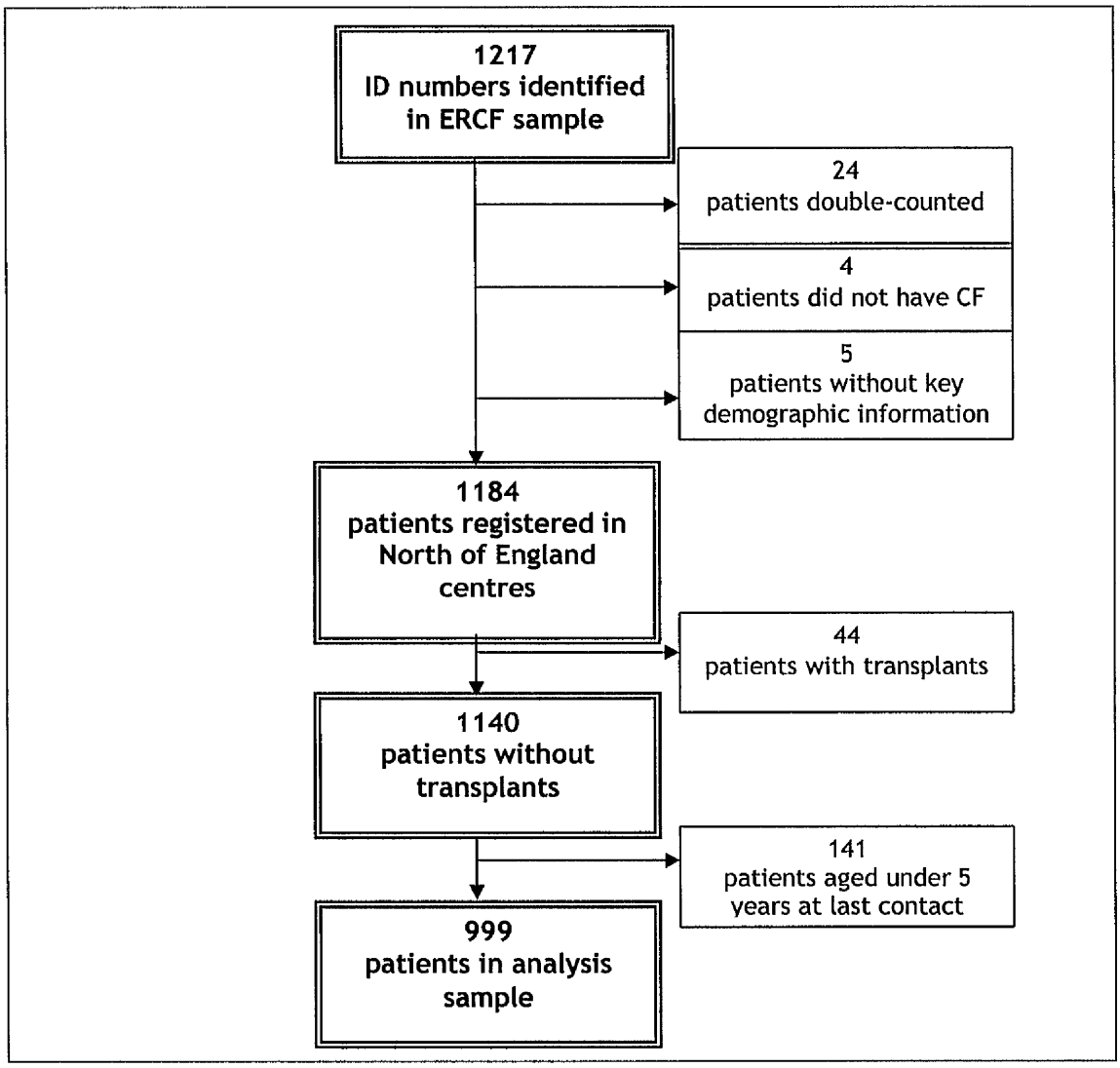
The DQR was much hampered by data protection legislation, particularly its lack of clarity; the required proceedings may have limited the effectiveness and certainly the efficiency of the DQR.

Following this detailed examination of data quality within the ERCF, the next two sections describe the sample selected for analysis and explore whether this could be seen as representative of any meaningful target population to which any results might be generalisable.

## 4.4 Study sample selection

There were 1,217 patients reported on in the ERCF sample; 24 patients were double entered, having initially been registered in a paediatric centre and then again when transferred to an adult centre without the records being linked. Four non-CF patients and five records with missing key demographic information were excluded. For the Case Study analysis, 44 patients with transplants had to be excluded, resulting in a sample of 1,140 patients (Figure 4.2). After excluding under-five-year-olds, the final sample size was 999.

**Figure 4.2: Selection of study sample**



## 4.5 Representativeness

### 4.5.1 Introduction

Registries and studies based on their data are often considered to represent the clinical realities better than randomised controlled trials (Rittenhouse & O'Brien 1996). In addition, a broader spectrum of the potential patient population can be represented in a registry, whereas RCTs (and thus their findings) are limited to patients carefully selected according to narrowly defined inclusion and exclusion criteria. However, representativeness of a registry cannot be assumed, and neither can generalisability of a registry-based study (Lewis 2001).

The US General Accounting Office (GAO 1992) identified two major categories of work assessing the generalisability of existing studies. The first assesses the *intention* in the patient selection process to recruit a representative patient sample, whereas the second focuses on the *results* of the patient selection process, i.e. whether a representative sample has been achieved, judged by empirical comparisons with patient pools seen in medical practice.

This section records an attempt to determine the generalisability of the Case Study by eliciting the representativeness of the patient sample. This requires that the sample and target population are similar in relevant characteristics. As target populations I used both the UK CF population, and the European ERCF population, and the two following sections present comparisons of the dataset with those populations respectively. For the UK population there were relatively few comparisons possible as published data were limited.

### 4.5.2 UK prevalence data

The prevalence of CF in the UK in the year 2000 has been estimated to be 7,750 cases, an average annual rise of about 2.2% based on an estimated 6,500 cases in 1992 (Dodge *et al.* 1997). The "North of England" dataset from the seven participating centres contains contacts with between 754 and 932 patients a year during its six years of data collection (see Table 4.9). This means that between 11% and 13% of the UK patients were in contact with the Northern centres in any one year.

Table 4.9 also illustrates that in 1994, the year of setting up the registry, the majority of patients were already recruited, but substantial numbers were added in the following years. After 1996, one centre no longer contributed data, and by 1999 further recruitment and reporting began to cease, which is reflected in declining numbers of patients reported on.

**Table 4.9: Number of patients reported on during each calendar year**

| Year | Number of patients reported on (*n*=1184) |
|---|---|
| 1994 | 754 |
| 1995 | 910 |
| 1996 | 932 |
| 1997 | 846 |
| 1998 | 864 |
| 1999 | 808 |

The UK Cystic Fibrosis Survey (UKCFS) claims to hold data on all people resident in the UK with CF born since 1968 and before that year, if they were still alive in 1977 (Dodge *et al.* 1997). Table 4.10 shows the number of cases reported in the UKCFS (Dodge *et al.* 1997) by their year of birth compared with the patients from the North of England dataset born during the same years (n=975). The per cent of UK cases registered in that dataset varied between 7% and 18% after 1969, with a rising trend towards later years. This may be because more patients from earlier birth years have already died

before 1994 when the ERCF began enrolling patients. This comparison confirms that the North of England dataset contained data on about 12% of all known CF patients in the UK. However, complete coverage of any particular geographic area is very unlikely, even for centres with complete registration, as centres do not draw patients only from clearly demarcated regions, and not all shared care patients are registered.

Knowing that our dataset covered a reasonable proportion of the national population does not by itself guarantee generalisability of any findings. No further comparisons of the UK population were possible; it is known for example that some genotypes are associated with more severe forms of the disease, and that genotypes vary geographically, probably also within this country. Similarly, the population coverage by specialist centres as well as treatment practices will vary - neither of which is sufficiently well described nationally.

**Table 4.10: Year of birth for North of England and UK CF patients and per cent of UK cases registered in North of England dataset (975 patients born 1968 to 1994)**

| Year of birth | Number of patients in North of England dataset | Cases reported in UKCFS (Dodge *et al.* 1997)* | North of England patients as % UKCFS cases |
|---|---|---|---|
| 1968 | 16 | 367 | 4% |
| 1969 | 17 | 356 | 5% |
| 1970 | 38 | 388 | 10% |
| 1971 | 42 | 407 | 10% |
| 1972 | 23 | 326 | 7% |
| 1973 | 32 | 309 | 10% |
| 1974 | 35 | 316 | 11% |
| 1975 | 31 | 291 | 11% |
| 1976 | 31 | 299 | 10% |
| 1977 | 32 | 278 | 12% |
| 1978 | 54 | 292 | 18% |
| 1979 | 51 | 302 | 17% |
| 1980 | 52 | 302 | 17% |
| 1981 | 44 | 302 | 15% |
| 1982 | 38 | 356 | 11% |
| 1983 | 35 | 301 | 12% |
| 1984 | 22 | 310 | 7% |
| 1985 | 46 | 311 | 15% |
| 1986 | 36 | 276 | 13% |
| 1987 | 38 | 294 | 13% |
| 1988 | 48 | 326 | 15% |
| 1989 | 42 | 290 | 14% |
| 1990 | 30 | 295 | 10% |
| 1991 | 41 | 259 | 16% |
| 1992 | 37 | 239 | 15% |
| 1993 | 33 | 202 | 16% |
| 1994 | 31 | 205 | 15% |
| TOTAL | 975 | 8199 | 12% |

*UKCFS=UK Cystic Fibrosis Survey

## 4.5.3 Comparison of the North of England with published ERCF data

This section compares the sample with ERCF data previously published and thus attempts to determine the representativeness of this European population. However, the coverage of the European CF population by the ERCF is in question; hence, the target population itself is a somewhat artificial creation. Comparisons were possible for several available variables: age, lung function baseline, treatment practices, dornase alfa use, and microbial colonisation.

### *4.5.3.1 Enrolment data: age and lung function*

Koch *et al.* (1997) published enrolment details of patients enrolled on the ERCF between January 1994 and December 1995. Of the patients from the seven "North of England" centres participating in the current study, 79% (931 patients) were enrolled during that period. A comparison of this subset with the enrolment data published by Koch and colleagues is presented below.

Table 4.11 compares age and lung function data on enrolment from the two sources. Table 4.12 compares data from the North of England sample (n=1184) and enrolment data from the ERCF for the period 1994 to 1995 (Koch *et al.* 1997) broken down by $FEV_1$ (%predicted) and age group. Comparisons between the ERCF population and the North of England sample are hampered firstly because of small numbers in the subgroups. Also, Koch *et al.* seemed to have excluded some 7% of registered patients because of missing data. It was unclear for example, whether missing lung function values on enrolment were a cause for exclusion; whereas no such missing data were indicated in any age groups, the age group under 6 years, where a significant number of missing values would be expected, still comprised 24% of their population. If missing data were excluded from our sample in these age groups, the proportions of the two samples would drift further apart to demonstrate even more starkly

the older age of the current study sample. However, it is unknown how many children's and adult centres respectively contributed to the ERCF at the period in question.

**Table 4.11: Age and lung function at enrolment, 931 study patients enrolled 1994-1995, compared with ERCF data**

| | North of England sample (95% confidence interval) | ERCF (*n*=6,858) (Koch *et al.* 1997) |
|---|---|---|
| Age (years) | n=931 | |
| Mean | 14.7 (14.1 - 15.3) | 13.6 |
| Median | 14.0 | 12.0 |
| FVC (% predicted) | n=723 | |
| Mean | 75.0 (73.3 - 76.6) | 78.0 |
| Median | 76.7 | 80.3 |
| $FEV_1$ (% predicted) | n=723 | |
| Mean | 64.2 (62.3 - 66.0) | 65.5 |
| Median | 65.4 | 66.5 |

**Table 4.12: Breakdown of age and lung function at enrolment, 931 study patients enrolled 1994-1995, compared with ERCF data**

| | | North of England sample | | ERCF (Koch *et al.* 1997) | |
|---|---|---|---|---|---|
| Age (years) | $FEV_1$ (% of predicted) | Patient number | % | Patient number | % |
| <6 | | 200 | 21.5 | 1,662 | 24.2 |
| 6-12 | <40 | 17 | 1.8 | 127 | 1.9 |
| | 40-70 | 52 | 5.6 | 504 | 7.3 |
| | 71-100 | 95 | 10.2 | 913 | 13.3 |
| | >100 | 33 | 3.5 | 275 | 4.0 |
| | Missing | 23 | 2.5 | | |
| 13-17 | <40 | 10 | 1.1 | 219 | 3.2 |
| | 40-70 | 52 | 5.6 | 461 | 6.7 |
| | 71-100 | 75 | 8.1 | 464 | 6.8 |
| | >100 | 13 | 1.4 | 80 | 1.1 |
| | Missing | 14 | 1.5 | | |
| 18 and over | <40 | 130 | 14.0 | 741 | 10.8 |
| | 40-70 | 145 | 15.6 | 841 | 12.3 |
| | 71-100 | 64 | 6.9 | 494 | 7.2 |
| | >100 | 5 | 0.5 | 77 | 1.1 |
| | Missing | 3 | 0.3 | | |
| Total | | 931 | 100.0 | 6,858 | 100.0 |

Table 4.11 and Table 4.12 raise the possibility that, at least in the first two years of the ERCF's existence, the North of England subset contained a slightly higher proportion of older patients than the ERCF as a whole. This may explain differences between the two sources in clinical variables such as lung function or also colonisation rates.

Table 4.13 and Table 4.14 present the mean and median $FEV_1$ and FVC % of predicted at enrolment. By comparison, the ERCF in 1998 reported mean values for $FEV_1$ of 90.5%, 78.9% and 67.3%, and 53.7% for the same age groups (<6 years, 6-12 years, 13-17 years, 18+ years respectively); the corresponding values for FVC were: 97.2%, 86.6%, 79.9%, and 71.4% respectively (Hodson *et al.* 1998a). However, these seem to be data based on multiple observations per individual, i.e. not only enrolment data.

**Table 4.13: $FEV_1$ on enrolment report by age at enrolment, 878 patients with enrolment $FEV_1$ available (North of England)**

| Age (years) | Number of patients | Mean $FEV_1$ (% of predicted) |
|---|---|---|
| <6 | 36 | 79 |
| 6-12 | 229 | 78 |
| 13-17 | 185 | 72 |
| 18 and over | 428 | 51 |
| Total | 878 | 64 |

**Table 4.14: FVC on enrolment report by age at enrolment, 877 patients with enrolment FVC available (North of England)**

| Age (years) | Number of patients | Mean FVC (% of predicted) |
|---|---|---|
| <6 | 36 | 78 |
| 6-12 | 229 | 83 |
| 13-17 | 185 | 84 |
| 18 and over | 427 | 66 |
| Total | 877 | 75 |

It appears that the North of England sample may have included somewhat more severely affected young children and adults, but adolescents showed better values than were reported in the European comparison data. The North of England data also compared less favourably with data from the US from 1990 (FitzSimmons 1994) which reported mean $FEV_1$ and FVC % of predicted of 68.9% and 80.9% respectively (mean and median age: 14.5 and 12.5 years respectively), compared to 64% and 75% in the North of England sample. However, at least some of that difference may be accounted for by the use of different prediction formulae (FitzSimmons used a modification of the Knudson formulae).

These findings not only demonstrate age differences between the sample and ERCF population, but also strongly suggest that there were differences in severity of each group, which could mean that the population of young children registered on the ERCF is comparatively well. However, given what we knew about the completeness of registration within our children's centres, this raises the question of whether the ERCF itself was representative of the actual European CF population. It is important to note that analysts of the ERCF population might not be aware of this problem or would not know the answer to the question, unless they truly know the completeness of registration from different centres (as we did from ours).

#### *4.5.3.2 Enrolment data: treatment practices*

There was also an opportunity to compare some treatment practices between the sample and the ERCF population. All centres reported using oral antibiotic prophylaxis routinely. Between 94% and 97% of patients reported on in each year had indications of continuous antibiotic prophylaxis recorded. Earlier reports on the ERCF in 1998 (Hodson *et al.* 1998a) indicate that 65.8% of patients were receiving continuous antibiotic prophylaxis; apart from centre (or indeed international) differences in routine clinical practice, the difference may also be due to a different definition being applied in the 1998

study (e.g. continuous prophylaxis indicated throughout the observation period). Nevertheless, this represents a significant difference in a key prophylactic intervention which could mean that both progression and outcomes of the disease might be inherently different between sub-populations of the ERCF.

Much of the comparison of treatment practices reported in the North of England sample and in the ERCF used data presented by Koch *et al.* (1997) (Table 4.15 to Table 4.20). Both sources indicate a rise in the use of oral bronchodilators with age and severity of pulmonary disease. In addition, inhaled bronchodilators saw an increased usage in patients with more severe pulmonary disease, but not age. In nearly all subgroups, the use of bronchodilators exceeded the number of patients recorded to be suffering from asthma-like symptoms. Overall, in the current study sample 36.2% of patients were reported to suffer bronchial hyper-reactivity or asthma-like symptoms on enrolment, in some subgroups 60% or more. The overall use of inhaled bronchodilators was reported to be 60%. By comparison Koch *et al.* reported a similar condition in up to 27% in any subgroup, whereas the indicated use of inhaled bronchodilators was in most age groups more similar to the rates observed by the current study, but about two or three times higher than the number of patients diagnosed in the ERCF (Koch *et al.* 1997) population.

Table 4.15 illustrates a point already made by Koch *et al.* (1997): their data indicated that in the UK inhaled corticosteroids are used far more (by 36% of patients on enrolment) than for example in France (10%) or Germany (12%). The authors put this down to the traditionally different approach to the treatment of asthma. The equivalent figure in the current sample is 34%. In both sources, the number of patients reported as being treated with oral steroids exceeds the number of patients reported to suffer from allergic bronchopulmonary aspergillosis (ABPA).

**Table 4.15: Inhaled corticosteroid treatment reported on enrolment, by age and lung function, 931 study patients enrolled 1994-1995, compared with ERCF data**

| | | North of England sample | | ERCF (Koch *et al.* 1997) |
|---|---|---|---|---|
| Age (years) | $FEV_1$ (% of predicted) | Patient number | % | % |
| <6 | | 58 | 29 | 17 |
| 6-12 | <40 | 6 | 35 | 24 |
| | 40-70 | 27 | 52 | 26 |
| | 71-100 | 38 | 40 | 21 |
| | >100 | 7 | 21 | 26 |
| 13-17 | <40 | 7 | 70 | 28 |
| | 40-70 | 21 | 40 | 34 |
| | 71-100 | 28 | 37 | 26 |
| | >100 | 6 | 46 | 21 |
| 18 and over | <40 | 43 | 33 | 41 |
| | 40-70 | 47 | 32 | 31 |
| | 71-100 | 17 | 27 | 22 |
| | >100 | 0 | 0 | 17 |

**Table 4.16: Inhaled bronchodilator treatment reported on enrolment, by age and lung function, 931 study patients enrolled 1994-1995, compared with ERCF data**

| | | North of England sample | | ERCF (Koch *et al.* 1997) |
|---|---|---|---|---|
| Age (years) | $FEV_1$ (% of predicted) | Patient number | % | % |
| <6 | | 82 | 41 | 36 |
| 6-12 | <40 | 11 | 65 | 63 |
| | 40-70 | 32 | 62 | 57 |
| | 71-100 | 46 | 48 | 47 |
| | >100 | 13 | 39 | 48 |
| 13-17 | <40 | 10 | 100 | 74 |
| | 40-70 | 39 | 75 | 63 |
| | 71-100 | 49 | 65 | 53 |
| | >100 | 7 | 54 | 45 |
| 18 and over | <40 | 102 | 78 | 80 |
| | 40-70 | 106 | 73 | 68 |
| | 71-100 | 41 | 64 | 52 |
| | >100 | 1 | 20 | 46 |

**Table 4.17: Oral bronchodilator treatment reported on enrolment, by age and lung function, 931 study patients enrolled 1994-1995, compared with ERCF data**

| | | North of England sample | | ERCF (Koch *et al.* 1997) |
|---|---|---|---|---|
| Age (years) | $FEV_1$ (% of predicted) | Patient number | % | % |
| <6 | | 2 | 1 | 3 |
| 6-12 | <40 | 0 | 0 | 5 |
| | 40-70 | 2 | 4 | 7 |
| | 71-100 | 1 | 1 | 4 |
| | >100 | 1 | 3 | 5 |
| 13-17 | <40 | 1 | 10 | 13 |
| | 40-70 | 2 | 4 | 7 |
| | 71-100 | 5 | 7 | 6 |
| | >100 | 1 | 8 | 5 |
| 18 and over | <40 | 34 | 26 | 19 |
| | 40-70 | 34 | 23 | 15 |
| | 71-100 | 9 | 14 | 12 |
| | >100 | 1 | 20 | 10 |

**Table 4.18: Asthma-like symptoms reported on enrolment, by age and lung function, 931 study patients enrolled 1994-1995, compared with ERCF data**

| | | North of England sample | | ERCF (Koch *et al.* 1997) |
|---|---|---|---|---|
| Age (years) | $FEV_1$ (% of predicted) | Patient number | % | % |
| <6 | | 47 | 24 | 14 |
| 6-12 | <40 | 5 | 29 | 25 |
| | 40-70 | 17 | 33 | 22 |
| | 71-100 | 18 | 19 | 17 |
| | >100 | 4 | 12 | 11 |
| 13-17 | <40 | 6 | 60 | 27 |
| | 40-70 | 12 | 23 | 21 |
| | 71-100 | 23 | 31 | 14 |
| | >100 | 3 | 23 | 10 |
| 18 and over | <40 | 75 | 58 | 24 |
| | 40-70 | 90 | 62 | 21 |
| | 71-100 | 25 | 39 | 15 |
| | >100 | 1 | 20 | 13 |

The use of pancreatic enzymes seems broadly similar in both sources and is comparable with other published reports of the prevalence of pancreatic insufficiency in CF patients. Of the 1,140 patients, 1,085 (96%) have any use of pancreatic enzyme recorded. This figure is slightly higher than the 92% of screened children found to suffer pancreatic insufficiency at 1 year of age (Bronstein et al (1992) quoted in Littlewood and Wolfe (2000)). These authors also observed lower proportions of pancreatic enzyme users in higher age groups and more severe patients. Similarly, but broken down by age at enrolment, a clear trend is visible towards patients who are enrolled later in life (and conceivably of a less severe phenotype) being less likely to use pancreatic enzymes.

**Table 4.19: Pancreatic enzyme treatment reported on enrolment, by age and lung function, 931 study patients enrolled 1994-1995, compared with ERCF data**

| | | North of England sample | | ERCF (Koch *et al.* 1997) |
|---|---|---|---|---|
| Age (years) | $FEV_1$ (% of predicted) | Patient number | % | % |
| <6 | | 189 | 95 | 93 |
| 6-12 | <40 | 16 | 94 | 97 |
| | 40-70 | 50 | 96 | 95 |
| | 71-100 | 90 | 95 | 94 |
| | >100 | 33 | 100 | 94 |
| 13-17 | <40 | 10 | 100 | 94 |
| | 40-70 | 50 | 96 | 94 |
| | 71-100 | 71 | 95 | 93 |
| | >100 | 12 | 92 | 86 |
| 18 and over | <40 | 113 | 87 | 90 |
| | 40-70 | 124 | 86 | 87 |
| | 71-100 | 52 | 81 | 84 |
| | >100 | 4 | 80 | 78 |

### *4.5.3.3 Dornase alfa use*

In Koch's *et al.* (1997) data from 1994 and 1995 the rates of dornase alfa treatment reported on enrolment varied between European countries from 16% in the UK, 25% in Germany, to 49% in France, giving an overall rate of 25%. The equivalent figure for the North of England dataset was 14%. Table 4.20 seems to indicate that the North of England dataset contained more users amongst children but fewer adults on dornase alfa compared to the overall UK figures.

This is an important finding as it demonstrates considerable differences in treatment decisions (which are likely to be affected at least as much if not more by economic than clinical factors). This adds further complexity to any analysis of the likely effect of dornase alfa based on this database.

**Table 4.20: Dornase alfa treatment reported on enrolment, by age and lung function, 931 study patients enrolled 1994-1995, compared with ERCF data**

| | | North of England sample | | ERCF (Koch *et al.* 1997) | |
|---|---|---|---|---|---|
| | | | | UK subset of ERCF | Total ERCF |
| Age (years) | $FEV_1$ (% of predicted) | Patient number | % | % | % |
| <6 | | 5 | 3 | 2 | 5 |
| 6-12 | <40 | 8 | 47 | 40 | 50 |
| | 40-70 | 14 | 27 | 21 | 38 |
| | 71-100 | 10 | 11 | 6 | 17 |
| | >100 | 2 | 6 | 5 | 13 |
| 13-17 | <40 | 7 | 70 | 54 | 57 |
| | 40-70 | 21 | 40 | 34 | 44 |
| | 71-100 | 5 | 7 | 9 | 21 |
| | >100 | 1 | 8 | 5 | 9 |
| 18 and over | <40 | 33 | 25 | 45 | 53 |
| | 40-70 | 13 | 9 | 21 | 34 |
| | 71-100 | 4 | 6 | 7 | 17 |
| | >100 | 1 | 20 | 0 | 5 |

### *4.5.3.4 Microbiological colonisation (period prevalence)*

Comparative data on microbiological colonisation of ERCF registered patients existed in the form of an abstract of a conference presentation in 2000 (Mastella *et al.* 2000). The authors selected 12,066 patients registered between January 1994 and August 1999 who had at least two valid cultures. It was unclear what was meant by "valid cultures" and also where exactly the included patients were from, but it seemed reasonable to assume that a substantial proportion of them were UK patients, since the UK was one of the largest contributors of data to the ERCF. Their data are presented in Table 4.21. The same table contains data from the North of England subset of 1,018 patients who had at least two culture tests recorded during the entire observation period (up to December 1999). No adjustments for different observation periods have been made; hence, in theory one may expect to see a slightly higher prevalence in the North of England sample.

The data seem to indicate a continuously higher prevalence of *Pseudomonas aeruginosa* across all age groups in the North of England centres, which was also higher than the overall prevalence of 83% in adolescents and adults reported by Penketh et al (1987).

For other micro-organisms (Table 4.22) the data also varied between the two sources (again, for the North of England sample only patients with at least two culture dates were included). Whereas for children and adolescents the reported prevalence of *B. cepacia* was very similar to the ERCF report, the prevalence in adults was 2.5 times higher in the North of England sample. This reflects the epidemic of *B. cepacia* infection in some adult centres in the early 1990s. Reported colonisation with *H. influenzae* was generally somewhat lower in the North of England sample, apart from that amongst adults, and reported colonisation with *S. aureus* was considerably lower, especially in children. Overall colonisation with the *H. influenzae* and *S. aureus* has previously been reported to be 68% and 60% respectively (Penketh *et al.* 1987). Colonisation with *Candida* shows higher rates in children and lower rates in adults of the North of England sample, and a generally more even distribution across age groups than that observed in the European data.

**Table 4.21: Patients with Pseudomonas aeruginosa colonisation reported during the observation period 1994-1999 by age group, 1,081 patients with more than one culture date recorded, compared with ERCF data**

| | North of England sample | | | ERCF data (Mastella *et al.* 2000) | | |
|---|---|---|---|---|---|---|
| Age (years) | N | Patients affected | % affected of all patients | N | Patients affected | % affected of all patients |
| <6 | 236 | 108 | 45.8 | 1894 | 720 | 38.0 |
| 6<13 | 358 | 234 | 65.4 | 3754 | 2109 | 56.2 |
| 13<18 | 286 | 238 | 83.2 | 2192 | 2066 | 70.9 |
| 18 and over | 486 | 457 | 94.0 | 4336 | 3600 | 83.0 |

Notes
Individual patients may appear in more than one cell.
All differences are significant at the 5% level (Chi square test).

**Table 4.22: Patients with any reported colonisation by different bacteria during the observation period 1994-1999 by age group, 1,081 patients with more than one culture date recorded, compared with ERCF data (Mastella *et al.* 2000)**

| | Staph. aureus | | B. cepacia | | H. influenzae | | Candida | |
|---|---|---|---|---|---|---|---|---|
| | NE | ERCF | NE | ERCF | NE | ERCF | NE | ERCF |
| Age (years) | % | % | % | % | % | % | % | % |
| <6 | 18.2 | 43.7 | **1.3** | **1.5** | 35.2 | 45.6 | 36.4 | 19.8 |
| 6<13 | 36.9 | 60.0 | **3.6** | **3.9** | 30.4 | 40.7 | **31.0** | **29.9** |
| 13<18 | 34.3 | 57.4 | **7.3** | **7.2** | **22.7** | **25.2** | 24.5 | 32.2 |
| 18 and over | 42.2 | 54.5 | 24.9 | 10.2 | 27.0 | 23.0 | 27.0 | 40.1 |

Notes:
NE=North of England sample; ERCF data from Mastella et al. (2000)
Individual patients may appear in more than one age and bacterial group.
Only bold data pairs are not significantly different at the 5% level.
Same "N" base group per age group as in Table 4.21.

## 4.5.4 Concluding remarks

There was a small amount of published data on the UK CF population, which would have been the more interesting target population to choose for determining representativeness. However, very little detailed UK data had been published from the national CF surveys.

Therefore, much of the comparison relied on ERCF data. Whereas this still highlighted interesting parallels and differences, this approach was somewhat unsatisfactory, because any un-representativeness in the North England sample could have easily been repeated Europe-wide and would have been difficult to detect in that way. For example, the ERCF was supplied with data from CF centres, and patients not cared for in such centres are excluded - in all countries. Also, it is known that the UK was one of the largest contributors to the ERCF and it is therefore unlikely that the ERCF was representative of the European CF population. In addition, reporting errors could be similar

across the ERCF as the same data collection mechanisms and instruments were used.

My familiarity with the actual situation in our regional participating centres was key to interpreting some of the findings (e.g. the high *B. Cepacia* prevalence), but more importantly to raising questions about the completeness and representativeness of the ERCF itself (e.g. of younger age groups). Without such familiarity with the source and background of the data, analysts of the ERCF might find it difficult to interpret their findings. This issue highlights that determining representativeness and data verification are closely related activities in practice and are vital in the preparatory stages of ESDs. Without confidence in the data source, any differences between that and a possible target population cannot be interpreted, and generalisability will remain an elusive concept - even though it is held up as a strength of ESDs.

A range of published survey results offered further sources of comparisons. The usefulness of such comparisons is limited, if their own representativeness of any target population is unknown. Nevertheless, areas of uncertainty may be highlighted. As epidemics and outbreaks make it difficult to compare microbiological data across geographic regions, such data are less useful in determining representativeness. A similar argument may apply to genotype information, which shows natural geographic variations. This makes the interpretation of any found variations difficult, as it remains unclear how much of the variation is due to true geographic difference, and how much to sampling error.[2]

[2] For example, in 81% of the sample (n=1,140), at least one DF508 mutation was reported. This is higher than the figure quoted by Rosenstein (2000) (66%) for the Caucasian population. In addition, 6% have at least one G551D mutation recorded, and 1.6% show a R117H mutation. Again, the prevalence of these mutations has been indicated to be several times lower in other populations (1.6% and 0.3% respectively) (Rosenstein 2000).

## 4.6 Analysis

The detailed analysis of the dataset was undertaken by an analyst. I am therefore only able to present the analysis provided and critique it vis-à-vis my knowledge of the data and relevant literature (see next chapter). This small section describes the approach of the analyst and in particular his relevant concerns.

The "allocation" to dornase alfa, including pre-and post-registration status, was used to divide the study population into five groups for comparison. It should be remembered that little is known of why some patients do, and others do not, respond to dornase alfa. Any categorisation without randomisation may therefore be confounded by these unknown factors.

Five groups were produced from the application of two concepts:

(A) Continuity of use
(B) Use at time of enrolment

The resulting groups are presented in Table 4.23. As this study was based on historical data, the five groups with different allocation of treatment had to be taken as given, and interpretation based on a review of data quality, and the recorded characteristics of the patients and changes in variables and confounders over time.

**Table 4.23: Groups analysed by dornase alfa use**

| Group | Use of dornase alfa | | | | | Total patients |
|---|---|---|---|---|---|---|
| | Continuous | | Some | | None | |
| | *Starting before enrolment* | *Starting on or after enrolment* | *Starting before enrolment* | *Starting on or after enrolment* | | |
| Group ID number | 1 | 2 | 3 | 4 | 5 | |
| After excluding transplants | 106 | 195 | 41 | 107 | 691 | 1140 |
| Further exclusion of under-5-year olds | 106 | 195 | 41 | 105 | 552 | 999 |

The principal outcome variables were death, exacerbation, lung function change, and anthropometric measures. Data at enrolment were reviewed for ethnic group, sex, age, age at diagnosis, diabetic status (and age at diagnosis), $FEV_1$ and FVC, height, weight and body mass index (BMI).

In the investigation of lung function and anthropometric data, reports were amalgamated if they fell into the same calendar quarter. The advantages of this approach were that arbitrary selection was avoided and all available data used. The disadvantage is that marked short-term changes would be averaged out. Given the length of the study period, this was considered a practical approach.

A particular problem in measuring lung-function decline - even in non-CF populations - is to identify if it is a true ageing effect. The period of observation (although far longer than most published CF studies on dornase alfa) allowed only limited comparison with general lung-function data. Testing for the significance of differences is complex as the data have been standardised, and it was unclear whether the quantity or quality of the data would allow the use of specialised tests. The current limited objective was to study annualised $FEV_1$ and FVC change rates, compiled from comparing the

averages for readings in the first and last quarters of observation, and dividing for the period of observation. Calculations were made for both the Dundee and Knudson formulae. A subsidiary analysis was made of the group with no known use of dornase alfa prior to registration (Group 2).

Data display was intended to show the main effects apparent in the sample and to facilitate comparison with the literature. The analyst decided that the testing of hypotheses in the study sample would present difficulties for the following reasons:

1. Without random allocation to groups, there can be no assurance that unrecorded factors "cancel out". For example, little information was available on why, at the time of enrolment, some patients were on dornase alfa, and others were not. Therefore, any test between groups for a particular factor may be undermined by confounding influences.
2. Lung function and anthropometric measures are conventionally standardised (percentage predicted lung function, height centile). It is not well appreciated that such procedures often invalidate the use of statistical procedures (e.g. t-tests) which assume particular distributions, because the original measurement unit has been discarded. Although other families of tests exist, these may lack power and require much larger numbers of patients.
3. Where repeated observations are made over time on the same individuals, the measurements are not independent, may not be at regular intervals or for the same overall time, and irregularities of timing or missing data may not be random (e.g. particularly ill patients may have more readings, or neglect routine appointments). All these features make analysis difficult.
4. Technical issues exist on separating out the influence of age, cohort, and period. Recent advances in treatment of CF (notably nutrition and antibiotics) have considerably increased life expectancy, and patients

will have had a wide variety of treatments in childhood. In this study, data were reviewed both by cross-sectional age groups (e.g. on enrolment) and by following a cohort in time.

The following chapter summarises the findings. It must be appreciated - as indicated by the analyst's approach - that the analysis is limited to descriptive information to facilitate comparisons with the literature, and does not enable any conclusions about the effect of dornase alfa treatment.

# 5 Dornase Alfa Case Study: Findings

## 5.1 Results

The following quote summarises well the complexities of research on CF:

> *"This [CF] patient population has proved challenging to clinical investigators for several reasons, including (1) a large intrasubject and intersubject variability in commonly used outcome measures (e.g. pulmonary function testing), (2) the progressive nature of the illness, resulting in strong period effects, (3) wide variation in illness severity, (4) intercenter variability in treatment regimens, (5) confounding effects of multiorgan dysfunction on pulmonary disease, (6) the occurrence of periodic exacerbations of pulmonary symptoms, and (7) the limited availability of research tools for measuring progression of disease in younger patients with mild pulmonary involvement."*
>
> (Ramsey & Boat 1994 p.178)

Due to the reasons described in the previous section, and not least the small sample size, the analysis is limited to descriptive approaches. This section of the report presents results in three blocks:

(1) The movement of patients into and out of the cohort (subsection 5.1.1)

(2) Baseline data examining the main variables, their relationships, and their distribution amongst the comparison groups (subsection 5.1.2), and

(3) Presentation of longitudinal observations on sample outcomes data (death, and lung function change) (subsections 5.1.3 and 5.1.4).

## 5.1.1 Enrolment, follow-up, and discontinuations

In this section, the patterns of enrolment, follow-up, and discontinuations of the entire sample of 1,184 patients are examined.

### *5.1.1.1 Enrolment*

The ERCF intended that participating centres would enrol all their patients onto the database. However, one children's centre had missed an estimated 10 patients, and one adult centre reported that it stopped enrolling patients for the ERCF after a certain number had been reached (about half to one third of the actual patient population of that centre). A further children's centre only contributed data for the first two years after which its participation in the ERCF ceased. Whereas for most of this centre's patients the follow-up period was thus relatively short, some of them have been followed up in one of the adult centres.

Most patients were enrolled into the ERCF during 1994, the year it was initiated. In fact, enrolment slumped during 1996-8; Centre 2 had stopped recruiting patients to the ERCF, and Centre 7 was not enrolling any new patients until 1999 (Table 5.1). Only Centre 6 - and possibly Centre 1 - exhibit a steady influx of patients over the years following 1994.

**Table 5.1: Number of patients enrolled in each centre per calendar year**

| Centre | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | Total |
|---|---|---|---|---|---|---|---|
| 1 | 115 | 13 | 7 | 6 | 6 | 13 | 160 |
| 2* | 113 | 18 | 6 | | | | 137 |
| 3 | 100 | 42 | 10 | 17 | 22 | 8 | 199 |
| 4 | 84 | 59 | 5 | 8 | 1 | 24 | 181 |
| 5 | 52 | 15 | 11 | 10 | 2 | | 90 |
| 6 | 165 | 13 | 13 | 15 | 14 | 12 | 232 |
| 7 | 122 | 20 | 13 | | | 29 | 184 |
| Total | 751 | 180 | 65 | 56 | 45 | 86 | 1183 |

* This centre stopped participating in the ERCF during 1996.

#### *5.1.1.2 Deaths, discontinuations, and transfers*

Ninety-nine patients (8.4%) included in the sample were reported to have died during the observation period (up to 1999). The mean and median ages at death reported were 23.9 (SD: 7.1) and 23.4 years respectively. Of all deaths, 83% were caused by cardio-respiratory failure (3% unknown).

Sixteen patients were reported to have moved away, and another two were reported as lost to follow-up; one patient was reported to have requested to leave the database and another one was discontinued for undefined "administrative" reasons.

During the observation period, 159 patients (13%) were transferred, two of them twice. Most transfers occurred between adult and children's centres within the same conurbation.

#### *5.1.1.3 Follow-up*

Table 4.9 above presented the number of patients reported on during each year. Overall, the regularity of reports varied considerably between individual patients; a sizable number of patients were only reported on once a year, some less often. This could have implications for the analysis, particularly if the reason for non-reporting is not known.

Sixty-two per cent of patients in the sample have been followed-up for over 3 years. The median period of observation was 1,480 days (inter-quartile range (IQR): 1,017), and the maximum was 2,100 days (5.8 years).

### 5.1.2 Baseline data

Table 5.2 gives baseline data by dornase alfa group, for sex, ethnicity, age at enrolment, age at diagnosis, diabetes status, and age at first report. There were relatively few intermittent dornase users who started treatment prior to enrolment (group 3), but every group had a reasonable representation of males and females.

The group with no record of dornase alfa was markedly younger at enrolment than the others. There were higher rates of diabetes in the groups using dornase alfa, compared to never-users.

**Table 5.2: Baseline data: Sex, race, age at enrolment, age at diagnosis, genotype, and diabetes status**

| | | Use of dornase alfa | | | | | Total |
|---|---|---|---|---|---|---|---|
| | | Continuous | | Some | | None | |
| | Sex | Starting before enrolment | Starting on or after enrolment | Starting before enrolment | Starting on or after enrolment | | |
| Group number | | 1 | 2 | 3 | 4 | 5 | |
| Sex | m | 56 | 80 | 21 | 58 | 327 | 542 |
| | f | 50 | 115 | 20 | 47 | 225 | 457 |
| | all | 106 | 195 | 41 | 105 | 552 | 999 |
| Non-Caucasians | | 3 | 2 | - | - | 13 | 18 |
| Mean age at enrolment (years) | m | 17.0 | 18.1 | 17.7 | 21.1 | 13.6 | 15.6 |
| | f | 16.6 | 16.9 | 17.4 | 18.0 | 14.2 | 15.7 |
| | all | 16.8 | 17.4 | 17.5 | 19.7 | 13.9 | 15.7 |
| Mean age at diagnosis (years) | | 1.6 | 2.2 | 1.0 | 3.1 | 2.8 | 2.54 |
| Diabetic - n | | 23 | 33 | 18 | 20 | 53 | 147 |
| % | | 22% | 17% | 44% | 19% | 8% | 13% |

Table 5.3 gives baseline data for % predicted $FEV_1$, FVC, and percentile height, weight, and body mass index standardised to the Child Growth Foundation 1990 references. For % predicted $FEV_1$ and FVC lung function, the "no

recorded use" group had substantially better values. Amongst dornase alfa users, those having started use before enrolment appeared to have slightly lower values, but this was less clear for females. For FVC, female values may have been rather higher than males.

Compared with the Child Growth Foundation references, the whole sample (n=999) was about the 33$^{rd}$ percentile for height and weight, and the 39$^{th}$ percentile for BMI, with little sex difference. Those with a record of continuous use of dornase alfa were 9 height percentiles, 18 weight percentiles, and 20 BMI percentiles lower than those in the group with no record of dornase alfa use.

**Table 5.3: Baseline data: $FEV_1$, FVC at enrolment (% predicted by Knudson formulae), height, weight, and BMI (standardised to Child Growth Foundation references), by sex and dornase alfa use group**

| | | Use of dornase alfa | | | | | Total |
|---|---|---|---|---|---|---|---|
| | | Continuous | | Some | | None | |
| | Sex | Starting before enrolment | Starting on or after enrolment | Starting before enrolment | Starting on or after enrolment | | |
| Group number | | 1 | 2 | 3 | 4 | 5 | |
| $FEV_1$ - n | | 80 | 174 | 39 | 102 | 434 | 829 |
| | m | 51.7 | 54.7 | 46.7 | 56.2 | 77.8 | 67.7 |
| | f | 48.5 | 59.7 | 58.0 | 56.8 | 78.1 | 66.5 |
| | all | 50.0 | 57.7 | 52.5 | 56.5 | 77.9 | 67.1 |
| FVC - n | | 80 | 174 | 39 | 102 | 432 | 827 |
| | m | 67.1 | 67.3 | 57.3 | 68.0 | 82.7 | 75.9 |
| | f | 62.7 | 73.2 | 72.3 | 73.3 | 86.4 | 78.1 |
| | all | 64.8 | 70.8 | 65.0 | 70.4 | 84.2 | 76.9 |
| Height percentile | m | 27.9 | 30.8 | 25.8 | 32.2 | 34.8 | 32.9 |
| | f | 23.9 | 32.7 | 30.3 | 33.3 | 34.8 | 32.7 |
| | all | 26.0 | 31.9 | 28.0 | 32.7 | 34.8 | 32.8 |
| Weight percentile | m | 21.3 | 23.2 | 24.1 | 24.4 | 37.6 | 31.9 |
| | f | 18.1 | 29.5 | 25.3 | 31.2 | 38.7 | 32.8 |
| | all | 19.7 | 26.9 | 24.7 | 27.4 | 38.0 | 32.3 |
| BMI Percentile | m | 26.6 | 27.0 | 33.3 | 27.0 | 45.9 | 38.7 |
| | f | 24.6 | 33.7 | 31.0 | 36.4 | 45.8 | 38.8 |
| | all | 25.7 | 31.0 | 32.2 | 31.2 | 45.8 | 38.7 |

This baseline overview already demonstrates stark baseline differences between dornase alfa treatment groups. A more detailed examination of lung function and anthropometric variables follows.

### *5.1.2.1 Lung function*

Table 5.4 presents % predicted $FEV_1$ by sex, age, diabetes status, and also deaths and exacerbation rates. There was no statistically significant difference for the distribution of "% predicted $FEV_1$" by sex. Unsurprisingly, deaths were associated with low values. Annualised "exacerbation" rates are more difficult to interpret - the statistically significant difference ($p<0.001$)

between dichotomised high and low rates appeared to relate to males and not to females.

The association of $FEV_1$ bands with age was statistically significant at $p<0.001$, indicating the progressive nature of the disease. The association of low % predicted $FEV_1$ with diabetes ($p<0.001$) was common to both males and females.

Table 5.5 presents the same % predicted $FEV_1$ groups by dornase alfa use status, sex, and initial 5-year age bands. There were similar male and female patterns in the relation of dornase alfa usage class with % predicted $FEV_1$ classes. The highest proportion of <40% values were in the two "continuous use" groups, but about 10% of those with no recorded use had these low values (there may, of course, be medical or personal decisions for not using dornase alfa).

**Table 5.4: Categorized "%predicted" $FEV_1$ lung function data (Dundee formulae) on enrolment by sex, initial age, diabetes, death, and exacerbation rates (n=834 with initial lung function categorisation)**

| | | Sex | | Deaths | | Dichotomised exacerbation rates (>= 1 yr observation) n = 732 | |
|---|---|---|---|---|---|---|---|
| | %pred. $FEV_1$ | Male | Female | Survivor | Death | Low | High |
| m | <40 | | | 71 | 28 | 26 | 56 |
| | 40<70 | | | 135 | 10 | 51 | 71 |
| | >70 | | | 203 | 5 | 110 | 67 |
| | all | | | 409 | 43 | 187 | 194 |
| f | <40 | | | 53 | 23 | 27 | 39 |
| | 40<70 | | | 127 | 15 | 66 | 63 |
| | >70 | | | 159 | 5 | 82 | 74 |
| | all | | | 339 | 43 | 175 | 176 |
| all | <40 | 99 | 76 | 124 | 51 | 53 | 95 |
| | 40<70 | 145 | 142 | 262 | 25 | 117 | 134 |
| | >70 | 208 | 164 | 362 | 10 | 192 | 141 |
| | all | 452 | 382 | 748 | 86 | 362 | 370 |
| Statistical significance - Chi-squared test | | P=0.32 | | all p<0.001 m, f: p<0.001 | | all p<0.001 m: p<0.001, f: p=0.26 | |

| | | Age | | Diabetes | |
|---|---|---|---|---|---|
| | %pred. $FEV_1$ | <18 | 18+ | No | Yes |
| m | <40 | 20 | 79 | 70 | 29 |
| | 40<70 | 67 | 78 | 130 | 15 |
| | >70 | 160 | 48 | 189 | 19 |
| | all | 247 | 205 | 389 | 63 |
| f | <40 | 18 | 58 | 49 | 27 |
| | 40<70 | 69 | 73 | 112 | 30 |
| | >70 | 134 | 30 | 144 | 20 |
| | all | 221 | 161 | 305 | 77 |
| all | <40 | 38 | 137 | 119 | 56 |
| | 40<70 | 136 | 151 | 242 | 45 |
| | >70 | 294 | 78 | 333 | 39 |
| | all | 468 | 366 | 694 | 140 |
| Statistical significance - Chi-squared test | | all p<0.001 m, f: p<0.001 | | all p<0.001 m, f: p<0.001 | |

**Table 5.5: Categorized "%predicted" $FEV_1$ lung function data (Dundee formulae) on enrolment by sex, dornase alfa user group, and age (<18/18+) (n=834 with initial lung function categorisation)**

| | %pred. $FEV_1$ | Use of dornase alfa | | | | | |
|---|---|---|---|---|---|---|---|
| | | Continuous | | Some | | None | All |
| | | Starting before enrolment | Starting on or after enrolment | Starting before enrolment | Starting on or after enrolment | | |
| | | 1 | 2 | 3 | 4 | 5 | |
| m | <40 | 22 | 21 | 9 | 19 | 28 | 99 |
| | 40<70 | 19 | 29 | 4 | 21 | 72 | 145 |
| | >70 | 13 | 18 | 6 | 17 | 154 | 208 |
| | all | 54 | 68 | 19 | 57 | 254 | 452 |
| f | <40 | 18 | 27 | 3 | 13 | 15 | 76 |
| | 40<70 | 22 | 39 | 11 | 20 | 50 | 142 |
| | >70 | 6 | 33 | 6 | 13 | 106 | 164 |
| | all | 46 | 99 | 20 | 46 | 171 | 382 |
| all | <40 | 40 | 48 | 12 | 32 | 43 | 175 |
| | 40<70 | 41 | 68 | 15 | 41 | 122 | 287 |
| | >70 | 19 | 51 | 12 | 30 | 260 | 372 |
| | all | 100 | 167 | 39 | 103 | 425 | 834 |

| | %pred. FEV1 | Initial five-year age group (years) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 5<10 | 10<15 | 15<20 | 20<25 | 25<30 | 30<35 | 35+ | Total |
| m | <40 | 4 | 6 | 19 | 31 | 14 | 16 | 9 | 99 |
| | 40<70 | 24 | 18 | 38 | 40 | 13 | 9 | 3 | 145 |
| | >70 | 78 | 42 | 55 | 24 | 5 | 3 | 1 | 208 |
| | all | 106 | 66 | 112 | 95 | 32 | 28 | 13 | 452 |
| f | <40 | 4 | 9 | 13 | 21 | 16 | 8 | 5 | 76 |
| | 40<70 | 22 | 28 | 37 | 26 | 20 | 5 | 4 | 142 |
| | >70 | 44 | 53 | 45 | 12 | 8 | 2 | | 164 |
| | all | 70 | 90 | 95 | 59 | 44 | 15 | 9 | 382 |
| all | <40 | 8 | 15 | 32 | 52 | 30 | 24 | 14 | 175 |
| | 40<70 | 46 | 46 | 75 | 66 | 33 | 14 | 7 | 287 |
| | >70 | 122 | 95 | 100 | 36 | 13 | 5 | 1 | 372 |
| | all | 176 | 156 | 207 | 154 | 76 | 43 | 22 | 834 |

### *5.1.2.2 Anthropometric data*

Table 5.6 summarises height, weight and BMI percentiles standardised to Child Growth Foundation references by sex and 5-year age group. It should be noticed that some of the figures were based on very small numbers. This cross-sectional analysis indicates that in males there was no marked change over age in percentile for height but a steady decrease for weight, with,

naturally, decreasing BMI percentile. The female data differed - there appears to be no clear trend, and maybe a "dip" in the 10-15-year age group.

**Table 5.6: Percentile of height, weight, and BMI data on enrolment, by sex and age (standardised to Child Growth Foundation 1990 references)**

| Age (years) | Height | Weight | BMI | n |
|---|---|---|---|---|
| **Males** | | | | |
| 5 < 10 | 37.6 | 43.9 | 53.0 | 179 - 181 |
| 10 < 15 | 34.1 | 36.2 | 40.9 | 67 - 70 |
| 15 < 20 | 27.9 | 25.2 | 34.7 | 117 - 118 |
| 20 < 25 | 28.3 | 22.7 | 28.4 | 96 |
| 25 < 30 | 39.9 | 25.3 | 21.5 | 32 |
| 30 < 35 | 26.5 | 15.1 | 19.3 | 26 - 27 |
| 35 + | 35.5 | 24.8 | 23.0 | 13 |
| Total males | 32.9 | 31.9 | 38.7 | 531 - 537 |
| | | | | |
| **Females** | | | | |
| 5 < 10 | 31.1 | 36.9 | 47.3 | 126 - 127 |
| 10 < 15 | 25.4 | 26.0 | 35.0 | 89 - 91 |
| 15 < 20 | 33.0 | 31.5 | 40.4 | 102 |
| 20 < 25 | 36.8 | 31.4 | 31.5 | 58 - 59 |
| 25 < 30 | 41.4 | 37.7 | 33.3 | 44 |
| 30 < 35 | 38.8 | 39.3 | 36.0 | 16 |
| 35 + | 43.8 | 29.6 | 21.9 | 11 |
| Total females | 32.7 | 32.8 | 38.8 | 446 - 450 |

Table 5.7 summarises anthropometric data for diabetes, which is a potential confounder for some analyses. There were no differences between the sexes. The analysis by diabetes status showed much poorer percentiles for diabetic patients, but there is probably also an age effect at play here: those identified as diabetic on enrolment were young adults seven years older than the non-diabetic patients.

**Table 5.7: Baseline data: Height, weight, BMI percentiles (standardised to Child Growth Foundation references), by diabetic status**

| | Diabetes record | Height | Weight | BMI | n |
|---|---|---|---|---|---|
| Males | No | 33.7 | 33.7 | 40.8 | 465-472 |
| | Yes | 26.8 | 19.4 | 23.4 | 65 |
| Females | No | 33.5 | 34.0 | 40.3 | 367-371 |
| | Yes | 29.4 | 26.8 | 31.8 | 79 |
| All | No | 33.6 | 33.8 | 40.6 | 832-843 |
| | Yes | 28.2 | 23.4 | 28.0 | 144 |

In summary, the baseline data present an interesting picture of not entirely surprising associations between factors such as diabetes with poorer lung function on enrolment. Predictably also, dornase alfa is used more in patients with poorer lung function, bacterial colonisation with particularly Pseudomonas, and those with more severe genotypes (see Appendix A for more baseline data presentation). Thus, dornase alfa users were not comparable to non-users at baseline. Lung function data by age also pointed at the progressive nature of the disease.

The next sections of this chapter give an illustrative overview of the outcome variables death, and change in lung function in the different dornase alfa use groups. Further information relating to the outcomes of exacerbation frequency, and anthropometric changes is contained in Appendix A.

### 5.1.3 Outcome variable: Death

Numbers of deaths in the different comparison groups were small. Table 5.8 presents an analysis of death by sex, age, and dornase alfa use. The rate of death increased with age and is higher in females. The oldest group at entry had small numbers and would be highly selected by survivalship from a time of less advanced care. The no-use group had the lowest rate of death. This is hardly surprising given the strong relationship between dornase alfa use and poor lung function, diabetes, and not least age. Also, the analyst observed that dornase alfa was often initiated during episodes of exacerbation.

**Table 5.8: Death by sex, initial age group, and dornase alfa use (n=999)**

| | | Age (years) | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 <6 | 6 <11 | 11 <16 | 16 <21 | 21 <26 | 26 <31 | 31 <36 | 36+ | |
| **Male** | Alive | 88 | 106 | 80 | 103 | 67 | 32 | 16 | 9 | 501 |
| | Deaths | 0 | 1 | 1 | 16 | 11 | 3 | 7 | 1 | 40 |
| | All | 88 | 107 | 81 | 119 | 78 | 35 | 23 | 10 | 541 |
| **Female** | Alive | 64 | 79 | 86 | 82 | 52 | 30 | 11 | 9 | 413 |
| | Deaths | 1 | 4 | 8 | 14 | 7 | 7 | 2 | 1 | 44 |
| | All | 65 | 83 | 94 | 96 | 59 | 37 | 13 | 10 | 457 |
| **Total** | Alive | 152 | 185 | 166 | 185 | 119 | 62 | 27 | 18 | 914 |
| | Deaths | 1 | 5 | 9 | 30 | 18 | 10 | 9 | 2 | 84 |
| | All | 153 | 190 | 175 | 215 | 137 | 72 | 36 | 20 | 998 |
| | % deaths | 1% | 3% | 5% | 14% | 13% | 14% | 25% | 10% | 8% |

Note: Age at enrolment unknown for 1 patient.

| | | Use of dornase alfa | | | | | All |
|---|---|---|---|---|---|---|---|
| | | Continuous | | Some | | None | |
| | | Starting before enrolment | Starting on or after enrolment | Starting before enrolment | Starting on or after enrolment | | |
| **Group number** | | 1 | 2 | 3 | 4 | 5 | |
| **Males and females** | Alive | 86 | 167 | 32 | 89 | 540 | 914 |
| | Deaths | 20 | 28 | 9 | 16 | 12 | 85 |
| | All | 106 | 195 | 41 | 105 | 552 | 999 |
| | % Deaths | 19% | 14% | 22% | 15% | 2% | 9% |

## 5.1.4 Outcome variable: $FEV_1$ lung function change

### *5.1.4.1 Lung function change in the five/six-year cohort: annual data and estimates of mean annual change*

This section presents data for the cohort of patients enrolled in 1994 or 1995, who had been under observation for five or six years. Comparisons of $FEV_1$ lung function over time are presented for males and females by broad age group (<18 / 18+ years), and diabetes status (Table 5.9), and by dornase alfa usage group (Table 5.10). There is a clearer downward trend for women than men, particularly in the <18 age group, and non-diabetes group (Table 5.9).

**Table 5.9: Six-year cohort - $FEV_1$ "% predicted" (Knudson formulae) by sex, observation year, initial age (<18/18+), and diabetes status (n=571)**

| | | | Age | | Diabetes | |
|---|---|---|---|---|---|---|
| | | | <18 | 18+ | No | Yes |
| n - Sex | m | | 167 | 131 | 256 | 42 |
| | f | | 166 | 107 | 218 | 55 |
| | all | | 333 | 238 | 474 | 97 |
| Males | n - observations | | | | | |
| Year 1 | | 266 | 78.58 | 57.88 | 70.33 | 57.62 |
| Year 2 | | 286 | 79.83 | 60.11 | 72.05 | 57.84 |
| Year 3 | | 287 | 79.57 | 57.89 | 70.49 | 54.30 |
| Year 4 | | 293 | 79.22 | 58.09 | 69.48 | 53.82 |
| Year 5 | | 288 | 79.66 | 57.55 | 68.69 | 51.51 |
| Year 6 | | 209 | 81.25 | 56.86 | 67.21 | 50.97 |
| All yrs | | 1629 | 79.56 | 58.05 | 69.81 | 54.54 |
| Females | n -observations | | | | | |
| Year 1 | | 248 | 76.83 | 60.01 | 72.51 | 58.45 |
| Year 2 | | 264 | 75.71 | 59.83 | 71.51 | 58.28 |
| Year 3 | | 268 | 73.31 | 56.16 | 67.68 | 53.76 |
| Year 4 | | 262 | 71.36 | 54.67 | 65.01 | 52.06 |
| Year 5 | | 267 | 69.74 | 55.80 | 64.07 | 51.87 |
| Year 6 | | 182 | 66.90 | 53.40 | 60.84 | 50.28 |
| All yrs | | 1491 | 72.99 | 56.50 | 67.21 | 54.29 |
| Males & Females | n - observations | | | | | |
| Year 1 | | 514 | 77.69 | 58.84 | 71.33 | 58.10 |
| Year 2 | | 550 | 77.73 | 59.99 | 71.80 | 58.09 |
| Year 3 | | 555 | 76.44 | 57.08 | 69.18 | 53.99 |
| Year 4 | | 555 | 75.40 | 56.52 | 67.46 | 52.81 |
| Year 5 | | 555 | 74.72 | 56.72 | 66.54 | 51.71 |
| Year 6 | | 391 | 74.23 | 55.30 | 64.43 | 50.55 |
| All yrs | | 3120 | 76.28 | 57.34 | 68.62 | 54.40 |

Table 5.10 allows comparison of $FEV_1$ lung function by year across the five-dornase alfa use groups. Group 1 and 3 had relatively small numbers, but no gross imbalance by sex. The results are difficult to interpret. Over years 1 to 5 there was a general decrease in lung function, but higher for females than males (eight against three percentage points). The continuous user groups 1 and 2 showed this sex difference clearly. The differences between groups 1 and 2 might be due to selection or survival effects. Group 5 with no recorded dornase alfa use showed the same sex difference although lower losses over time.

**Table 5.10: Six-year cohort - $FEV_1$ "% predicted" (Knudson formulae) by sex, observation year, and dornase alfa groups (n=571)**

| | | Use of dornase alfa | | | | | All | n |
|---|---|---|---|---|---|---|---|---|
| | | Continuous | | Some | | None | | |
| | | Starting before enrolment | Starting on or after enrolment | Starting before enrolment | Starting on or after enrolment | | | |
| Group number | | 1 | 2 | 3 | 4 | 5 | | |
| Sex | m | 17 | 57 | 14 | 49 | 161 | | 298 |
| | f | 22 | 81 | 13 | 37 | 120 | | 273 |
| | all | 39 | 138 | 27 | 86 | 281 | | 571 |
| Males | | | | | | | | |
| Year 1 | | 60 | 58 | 53 | 59 | 78 | 68 | 266 |
| Year 2 | | 58 | 58 | 57 | 60 | 80 | 70 | 286 |
| Year 3 | | 56 | 55 | 57 | 58 | 78 | 68 | 287 |
| Year 4 | | 56 | 52 | 62 | 56 | 78 | 67 | 293 |
| Year 5 | | 53 | 52 | 60 | 54 | 77 | 66 | 288 |
| Year 6 | | 57 | 51 | 59 | 57 | 74 | 65 | 209 |
| All years | | 57 | 54 | 58 | 57 | 78 | 68 | 1629 |
| Females | | | | | | | | |
| Year 1 | | 58 | 65 | 57 | 62 | 79 | 70 | 248 |
| Year 2 | | 55 | 62 | 52 | 59 | 81 | 69 | 264 |
| Year 3 | | 51 | 57 | 49 | 57 | 77 | 65 | 268 |
| Year 4 | | 50 | 52 | 48 | 54 | 76 | 62 | 262 |
| Year 5 | | 49 | 51 | 47 | 53 | 75 | 62 | 267 |
| Year 6 | | 45 | 50 | 45 | 47 | 72 | 59 | 182 |
| All years | | 52 | 57 | 50 | 56 | 77 | 65 | 1491 |
| Males and females | | | | | | | | |
| Year 1 | | 59 | 62 | 55 | 60 | 79 | 69 | 514 |
| Year 2 | | 56 | 61 | 54 | 60 | 81 | 70 | 550 |
| Year 3 | | 53 | 56 | 53 | 57 | 78 | 67 | 555 |
| Year 4 | | 52 | 52 | 56 | 55 | 77 | 65 | 555 |
| Year 5 | | 51 | 52 | 54 | 54 | 76 | 64 | 555 |
| Year 6 | | 50 | 51 | 52 | 52 | 73 | 62 | 391 |
| All years | | 54 | 56 | 54 | 57 | 77 | 66 | 3120 |

Table 5.11 summarises for the cohort with 5 to 6 years of observation annualised $FEV_1$ change by age (<18 / 18+), diabetes status, dornase alfa group, and five-year age group at enrolment. The data comparison across dornase alfa use groups is difficult to interpret; women showed greater rates of decline than men. The decline in lung function was greater in younger age groups. The analysis by five-year age group at enrolment has reasonable group numbers only below age 25. There was no evidence of marked cohort

effects, but if the older groups represent less severe cases, the loss rates were rather greater than might be expected. Note that the male group aged 25<30 showed a *gain* of 1.6% per year, but with only 20 measurements this might be a random effect.

**Table 5.11: Six-year cohort - annualised change "% predicted" $FEV_1$ by sex, age (<18/18+), and diabetes status, use of dornase alfa, and initial age group (n= 473)**

| | | Age | | Diabetes | |
|---|---|---|---|---|---|
| | | <18 | 18+ | No | Yes |
| Sex | m | -.72 | -1.10 | -.71 | -2.10 |
| | f | -3.10 | -1.38 | -2.42 | -2.27 |
| | all | -1.95 | -1.22 | -1.50 | -2.19 |
| *n* | *m* | *125* | *121* | *211* | *35* |
| | *f* | *133* | *94* | *182* | *45* |
| | *all* | *258* | *215* | *393* | *80* |

| | | Use of dornase alfa | | | | | |
|---|---|---|---|---|---|---|---|
| | | Continuous | | Some | | None | All |
| | n | Starting before enrolment | Starting on or after enrolment | Starting before enrolment | Starting on or after enrolment | | |
| | | 1 | 2 | 3 | 4 | 5 | |
| m | 246 | -.82 | -1.45 | 1.88 | -1.49 | -.82 | -.91 |
| f | 227 | -2.30 | -3.19 | -2.74 | -2.42 | -1.79 | -2.39 |
| all | 473 | -1.61 | -2.52 | -.24 | -1.90 | -1.24 | -1.62 |

| | | Age group (years) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | n | 5<10 | 10<15 | 15<20 | 20<25 | 25<30 | 30<35 | 35+ | |
| m | 246 | -.76 | -.38 | -1.97 | -1.13 | 1.63 | -1.45 | -.88 | -.91 |
| f | 227 | -3.70 | -2.83 | -2.77 | -1.62 | -1.12 | -.68 | -.03 | -2.39 |
| all | 473 | -2.07 | -1.81 | -2.35 | -1.32 | .02 | -1.18 | -.43 | -1.62 |

### *5.1.4.2 "Before and after" dornase alfa initiation in one group*

Patients in Group 2 were enrolled with no recorded use of dornase alfa, but later commenced continuous use. A comparison pre- and post- initiation of dornase alfa was made to see if there were interpretable results.

Records for the entire group (n=195) were examined for the existence of $FEV_1$ lung function measures (based on Knudson formulae) for the quarter of original enrolment, the quarter prior to dornase alfa initiation, and a later period of observation. Only 101 sets of records were available to compute "before and after" rates; the numbers would be even lower if a minimum period of observation of one year was imposed.

**Table 5.12: Group 2 - lung function as "% predicted" $FEV_1$ (Knudson formulae) longitudinal comparison pre- and post dornase alfa initiation**

Rate computed on individual basis as percentage point change over time and summarized

| | | Division of observation period relative to dornase alfa initiation | |
|---|---|---|---|
| | | **Pre-initiation** | **Post-initiation** |
| Male N= 44 | Observed | 1.69 years | 2.29 years |
| | Annual rate | + 0.6 | - 2.1 |
| Female N= 57 | Observed | 1.54 years | 2.62 years |
| | Annual rate | - 1.3 | - 4.0 |
| All | Observed | 1.60 years | 2.48 years |
| | Annual rate | - 0.5 | - 3.2 |
| **Background data:** | | | |
| Age | 49 under 18, 52 18+ years | | |
| Deaths | 14 deaths, 87 survivors | | |
| Genotype | 92 with one or more DF508, 9 without | | |
| Diabetes | Children (<18) 5 of 49, adults (18+) 12 of 52 | | |

## 5.2 Discussion of results

The Dornase Alfa Case Study set out to analyse the experience of homogenous groups of dornase alfa users in terms of a number of outcome variables (death, respiratory exacerbation, lung function decline, and anthropometric changes). The results presented are descriptive and exploratory and intended to show differences between the groups, but also allow readers to compare the data with that from other samples. In this sample, dornase alfa users were clearly not comparable to non-users at baseline in terms of colonisation with *Pseudomonas* or *B. Cepacia*, initial lung function, anthropometric measures, diabetes, or severity indicated by genotype. Non-users were also younger at baseline. The sample also showed already familiar epidemiological patterns, with female patients being comparatively more severely affected.

Death was shown to be associated with age, except for the highest age group, which would be highly selected in terms of survivalship. Dornase alfa use groups showed considerably higher death rates than non-users. However, little can be said concerning the impact of dornase alfa on survival, as numbers of deaths in the sample were too small for meaningful comparisons between the dornase alfa use groups (Table 5.8), and the analysis does not permit any conclusions regarding the impact of dornase alfa due to the considerable baseline differences between treated and non-treated patients.

For less than 40% of deceased patients, lung function measurements from the last quarter of observation were available. For 26 males and 29 females, the $FEV_1$ at death was 29% and 33.5% respectively, i.e. close to the 30% level used in the literature to indicate a 50% chance of death in the next two years (Kerem *et al.* 1992). However, these are very small numbers, and there was wide scatter. For 43 patients for whom a rate of $FEV_1$ lung function decline

could be calculated, the rates for both males and females indicated much faster decline than in the patient population generally (-3.69% and -5.77% for males and females respectively).

The generally observed decrease in $FEV_1$ over time is higher for females than males (Table 5.10). Again, the differences between the non-user group (Group 5), and Groups 1 and 2 cannot be interpreted as Group 5 had much higher lung function, and this would affect selection for dornase alfa use.

The analysis by dornase alfa use pattern was equally difficult to interpret (Table 5.11). The patients with "no recorded dornase alfa use" lost $FEV_1$ more slowly than patients who started use after enrolment and have used it continuously, but how the users would have fared without it, or the non-users with it, is unknown; and lung function at commencement was different. As usual, female rates of decrease in lung function are higher. Group 2 (n=195) was examined in detail. Although the records indicated general observation of the UK licensing requirement for at least 40% predicted FVC before commencement, only 101 sets of records were available to compute "before and after" rates. By inspection, the loss of lung function was *faster* after initiation onto dornase alfa, for both males and females. However, the detailed statistics showed marked variation, and there were no controls for the group. The data analyst noticed, in reviewing all data for the 101 patients, that there were a number of cases where dornase alfa appeared to have been initiated after large fluctuations in $FEV_1$ measurements, and it is possible, therefore, that dornase alfa was selectively offered in an acute situation and continued thereafter. This shows that treatment decisions which are unrecorded in the database and therefore unable to be accounted for, can have potential significant impact on the results of an ESD and their interpretation.

Overall, the data presented here did not allow conclusions regarding the effectiveness of dornase alfa, but also showed no indication that treatment improved disease progression. The comparatively poor "outcomes" in the dornase use groups were not surprising given the differences at baseline.

Johnson *et al.* (1999) undertook an observational study to estimate the effect of dornase alfa by using data from the North American ESCF. Patients aged 6 years or more with >=40% of predicted $FEV_1$ who had been enrolled for at least 18 months were eligible for the analysis. A baseline spirometry from at least 6 months post-enrolment had to be available, as well as a test approximately 12 months after (treated patients were required to have started dornase alfa at least 6 months after enrolment). The primary outcome was change in $FEV_1$ (% predicted) from baseline. The multiple regression analysis retained 18 variables in the final model, after having considered 49. Those retained included the number of exacerbations prior to baseline spirometry, age, weight-for-age percentile, baseline $FEV_1$, wheeze, cough, sex, centre affiliation and type of practice, several micro-organisms detected at respiratory culture, and the use of inhaled corticosteroids, supplemental oxygen, and bronchodilators. From my experience, several of these variables would have been highly unreliable in the ERCF, which raises questions about their potential reliability in the US dataset. There was no obvious attempt in the Johnson paper to account for data quality in a similarly rigorous manner to what I had attempted to do in the Dornase Alfa Case study.

The 283 patients treated for an average of 329 days (SD 45) showed an adjusted benefit of dornase treatment of 4.3 (SE: 0.9) points of predicted $FEV_1$ over 2,382 untreated patients. Treated and untreated patients started with a mean $FEV_1$ of 76.1% (SD: 19.1) and 87.5% (SD: 19.4) predicted respectively at baseline. After twelve months, treated and untreated patients showed 80.0% (SD: 21.0) and 85.9% (SD: 20.2) predicted $FEV_1$. An intention to treat analysis included patients starting the treatment but not continuing for a year or

changing the regimen. The adjusted benefit of dornase alfa was an estimated difference in mean change from baseline of 3.2 (SE: 0.7) in favour of the dornase group.

There are several differences between this and the Dornase Alfa Case Study. Firstly, patients in the Johnson study had markedly better baseline $FEV_1$ (Table 5.13), and were younger than those in the Case Study. It must be remembered that the Case Study Group 2, which is the best comparison group, would have included a wide range of treatment durations.

**Table 5.13: Differences between patients included by Johnson et al. (1999) and Dornase Alfa Case Study patients (Group 2) on enrolment**

| | Johnson et al 1999 | | Dornase Alfa Case Study | |
|---|---|---|---|---|
| | Continuous treatment | No treatment | Group 2 | No treatment |
| Mean $FEV_1$ %pred. at enrolment | 76.1% | 87.5% | 57.7% | 77.8% |
| Mean age (years) | 14.1 | 13.9 | 17.4 | 13.9 |
| N | 283 | 2382 | 195 | 552 |

In the current study all treated groups lost $FEV_1$ over the 6-year observation period, and the untreated group lost $FEV_1$ more slowly than treated patients of group 2 (-1.24% and -2.52% respectively). It is possible that Johnson captured an initial positive response in the treatment group; this would tally with the observations of a 2-year RCT where treated patients had returned to their baseline $FEV_1$ after two years, despite significant improvements after one year (Quan *et al.* 2001). It is thus possible that the initial improvement is not maintained over several years, and that more severely affected patients respond less favourably to treatment (which has been observed elsewhere by Hodson *et al.* 2003). UK centres may be comparatively less likely to use dornase alfa (according to the ERCF, this is certainly the case compared to other European countries) and hence reserve treatment for a more severely

affected group of patients. Therefore, any analysis of a UK dataset might might not be comparable to other populations.

The use of multivariate analytical approaches appears tempting. However, the analyst refrained from further analysis of the database. Johnson et al. (1999) also bemoan the low $R^2$ value in their study, despite having included nearly 50 variables (which in itself may be considered questionable). Thus the extent to which further confounding may have influenced their results is in question. Significantly, they have also stopped short of using death as an outcome.

Poor data quality and completeness of other key variables, in particular exacerbations, further deter from multivariate analytical approaches. Significantly, no explicit rationale was available for initiation of treatment for patients commenced on dornase alfa, and there was good indication that the decisions to offer dornase alfa differed between centres. Recording practices in different centres invariably impacted on the data. Lung function values varied little between individual years, but did vary between centres (e.g. one centre reported that relatively "good" values were selected and recorded for each observation period between visits; this centre did show higher values compared to others).

It is interesting that Johnson *et al.* (1999) did not comment on data quality; they stated that "*patients were eligible for entry into the analysis data set based primarily on availability of data*" (p.735). The numbers of patients excluded were not reported, and neither was any comparison between included and excluded patients. Thus, the generalisability of the study is difficult to judge, regardless of how complete the database itself might have been in terms of capturing the North American CF population. Their sample may also be subject to further selection bias, as patients had to survive for the observation period to be included.

It is also interesting to observe that a later published analysis of the ERCF entitled "Dornase alfa in the treatment of cystic fibrosis: a report from the Epidemiologic Registry of Cystic Fibrosis" (Hodson *et al.* 2003) stopped short of multi-variate analysis, although the authors had access to the entire European database including 13,684 patients. Nevertheless, they only presented a descriptive un-controlled comparison of treated and un-treated patients over two years, examining change in $FEV_1$, and numbers of exacerbations (despite similar baseline differences to the ones reported here). Despite the baseline differences, the study reported a larger reduction in exacerbation rates in treated compared to non-treated patients (mean -0.19 and 0.06 respectively), as well as a slower reduction in lung function (2.5% and -1.1% change from baseline after one year, and 0.3% and -2.3% change from baseline after two years). Interestingly, no subsequent multi-variate analysis has ever been published, and my enquiries about this remained un-answered.

Hodson *et al.* (2003) as well as Johnson *et al.* (1999) differ in their conclusions (certainly in terms of lung function) from the results of the Dornase Alfa Case Study. A methodological study by Rothman and Wentworth (2003) described below provides valuable insights into the possible variations in results from multi-variate analyses in CF due to residual confounding, which might have affected the Johnson paper. Data quality aspects may indeed play a role in both papers, which we can only appreciate based on our detailed explorations prior to analysis. The UK has a low rate of dornase alfa use compared to other European countries, which may mean that patients receiving the drug here are more severely affected by the disease than the average European dornase users. Indeed, Hodson *et al.* (2003) report a mean $FEV_1$ %predicted of 60.2 for the treated group, whereas all dornase use groups in the Case Study started at lower mean baseline levels (between 50 and 57.7). There is some indication in the literature that less severely affected patients respond better to dornase alfa treatment.

The second outcome variable used by Hodson *et al.* (2003) is exacerbation frequency. Given the definitional problems with this variable identified in the Dornase Alfa Case Study, and the variable completeness in reporting between centres, as well as the fact that dornase alfa seems at least in some centres to be introduced in response to exacerbations, a reduction in exacerbations in the year following initiation seems perhaps less surprising.

The renowned statisticians Rothman and Wentworth (2003) skilfully illustrated the precarious nature of multivariate approaches in cystic fibrosis. The authors used US Cystic Fibrosis Foundation registry data in a retrospective cohort study of the risk of death of patients using or not using tobramycin. They found the crude risk of death to be 3.5 times higher in the treated group compared to the non-treated group. Controlling for age, lung function, height and infection with P. aeruginosa reduced the risk ratio to 1.2. The authors found it difficult to remove the strong confounding by indication. However, an analysis of patients who would have met the inclusion criteria of an RCT of tobramycin resulted in a risk ratio of 1.05 (or 1.06, depending on the duration of use). This vast change in risk ratio is taken as a strong indication that residual confounding could easily be masking a positive result in favour of tobramycin. It is worth remembering at this point, that the Case Study DQR identified 10% of deaths as missing from the registry!

The authors point out that sensitivity and specificity of confounder measures have a significant impact on residual confounding. Based on previous work by Savitz and Baron (quoted in Rothman and Wentworth, 2003) they assume that 75% of the confounding in the observed data might be controlled and calculate that the completely controlled risk ratio could theoretically be 0.46. This is close to the 0.43 (on protocol analysis) reported in the RCT. This study provides a convincing illustration that the interpretation of data greatly depends on assumptions and residual biases. It also is a stark reminder of how

vulnerable and difficult to interpret a multi-variate analysis of the Dornase Alfa Case Study would have been, given the existing concerns about the data quality of key confounders, such as exacerbations.

The Case Study analysis has paid little attention to the duration of dornase alfa use, and the pragmatic groupings hide a considerable variation of use patterns, as Table 5.14 demonstrates. Two thirds of dornase alfa users had started on or after enrolment (Groups 2 and 4), but a majority of Group 4 (intermittent users) used dornase alfa for less than one year. All other studies discussed here have only selected patients treated continuously over the study period. A selection of patients continuously treated over a significant length of time, with sufficient data available from the pre-initiation period, would have generated a small sample, too small for meaningful multi-variate analysis.

**Table 5.14: Cumulative duration of treatment of patients on dornase alfa (n=447) by use group**

| Total sum of days on treatment | Use of dornase alfa | | | | | | | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Continuous | | | | Some | | | | | |
| | Starting before enrolment | | Starting on or after enrolment | | Starting before enrolment | | Starting on or after enrolment | | | |
| 0 - 365 | 8 | 8% | 31 | 16% | 14 | 34% | 63 | 60% | 116 | 21% |
| 366 - 730 | 20 | 19% | 43 | 22% | 6 | 15% | 10 | 10% | 79 | 14% |
| 731 - 1095 | 18 | 17% | 37 | 19% | 7 | 17% | 12 | 11% | 74 | 13% |
| 1096 - 1460 | 9 | 9% | 36 | 19% | 5 | 12% | 11 | 11% | 61 | 11% |
| 1461 - 1825 | 39 | 37% | 36 | 19% | 8 | 20% | 7 | 7% | 90 | 16% |
| 1826+ | 12 | 11% | 2 | 1% | 1 | 2% | | | 15 | 3% |
| Unknown | | | 10 | 5% | | | 2 | 2% | 12 | 2% |
| Total | 106 | 100% | 195 | 100% | 41 | 100% | 105 | 100% | 447 | 100% |

This demonstrates that the study design for any ultimate ESD analysis should be foreseen and accounted for in the design of the database, as post-hoc

categorisations of comparison groups are fraught with difficulty, and poor definitions of key outcome variables cannot be remedied later.

## 5.3 Concluding remarks

The analysis of the Dornase Alfa Case Study confirmed some known epidemiological patterns in this sample of CF patients from the North of England: there was a slight predominance of males, which increased with higher age groups due to known survival advantages. Differences in genotype severity between conurbations were also observed. Diabetic patients tended to have lower lung function at baseline, and higher dornase alfa use rates. Not all observations were easily explained, and some may have arisen due to poor data quality.

The analysis also showed that there were variations in decisions to initiate dornase alfa treatment, which are neither clear nor recorded on the database. This means that an analyst of an ESD is largely unable to account for such variations albeit that they would have significant implications on group allocation and interpreting any findings. Similarly, there were differences in practice and recording between centres, resulting in systematic differences on certain variables.

As in other observational studies, the analysis showed a range of important baseline differences between patients treated and not treated with dornase alfa. Generally speaking, dornase alfa users were older, and less well. Treated patients had lower lung function values at baseline, and were at lower percentiles for height and weight, more likely to show microbiological colonisation, and more likely to be diabetic. Only a small number of the available variables have been used; however, it is already clear from the above that key information relevant for the analysis has not been collected in the database (e.g. reason for treatment decision, shared care, and compliance).

The analysis shows higher numbers of deaths, and a faster lung function decline and anthropometric deterioration in treated groups compared to non-treated patients. However, direct comparisons are difficult to interpret, as the large baseline differences would need to be accounted for. The groupings of dornase alfa users hide considerable differences in duration of treatment, which further impede interpretation.

A careful multi-variate analysis of a larger population, such as the entire UK ERCF population, might be of interest. However, given the likely residual confounding and poor quality of outcome variables (except possibly lung-function values), the results would remain very difficult to interpret. Hence, we abandoned further analysis attempts.

The experience of preparing and undertaking the Dornase Alfa Case Study thus had revealed a whole host of methodological concerns about the quality of data and their use in analysis. I was therefore interested to review whether other researchers having undertaken similar ESDs of drug treatments acknowledged similar problems, and how they might have addressed them. The next chapter describes my review of published ESDs in which I explore this question.

# 6 Comprehensive Review of ESDs of Drug Therapies

## 6.1 Introduction

The earlier literature review highlighted both the increasing popularity of as well as reservations about the use of databases in effectiveness research. Several systematic reviews of different aspects of comparisons of randomised and non-randomised studies have contributed to the methodological knowledge about study designs. The need for more methodological research has been highlighted, including a need for more direct evidence about the comparability of findings from different study designs (MacLehose *et al.* 2000). However, no systematic effort to assess ESDs and their contribution to effectiveness research has yet been undertaken.

ESDs follow non-randomised designs and much of the methodological evidence on those designs can be expected to be relevant to ESDs. However, the nature of the data sources used in ESDs may give rise to specific strengths and weaknesses. This review therefore aimed to identify ESDs, and to provide an overview of specific study features relating to their internal and external validity.

This work followed on from the Dornase Alfa Case Study which was essentially unable to answer the question of long-term effectiveness of dornase alfa. I hoped that a review of similar ESDs would confirm some of the methodological lessons highlighted in the Dornase Alfa Case Study, identify additional lessons beneficial to future research, and allow me to generalise and thus better describe some of the issues.

## 6.2 Aim

Explore in existing ESDs features relevant to their internal and external validity.

## 6.3 Objectives

1. Identify and retrieve published effectiveness evaluations of individual drug therapies using routinely available data or databases;

2. Describe ESDs by data sources and their use, and other features relevant to internal validity (including control of bias and confounding through study design and analysis, data quality, and results), and external validity (including sample selection and participation);

## 6.4 Methods

### 6.4.1 Search strategy

A gold standard systematic literature search not only examines several bibliographic databases, but also grey literature, and includes hand searching of journals. Whereas a systematic identification of all ESDs would have been desirable, it was clear from the outset that this was an unrealistic expectation, given the comparatively poor indexing of observational studies in bibliographical databases (MacLehose *et al.* 2000). These difficulties are compounded when databases are included as a search criterion. The intention thus was to follow the principles of a systematic literature search within existing resource constraints, and describe the methods used in sufficient detail to permit the reader to judge their limitations (see section 6.6). The exploratory nature of this work justifies the restrictions applied to the search methods. Due to these limitiations in identifying potential studies for inclusion, this review is described as "comprehensive" rather than "systematic".

Hence, the search strategy included two main groups of methodological terms: those relating to the data source, as well as those indicating effectiveness studies. The choice of free-text terms used was informed by the terminology used in previously identified relevant texts and articles (by having undertaken pilot searches). The search strategy (Table 6.1) was performed on Medline (1966-2002) on 24 October 2002. This review was limited to studies on drug therapies in humans.

Only journals included in the Abridged Index Medicus were searched (as indicated by the term "Core Clinical Journals" - see Table 6.1). This Index

comprises 120 high-quality core clinical English language journals (National Library of Medicine 2001). This restriction had been applied previously in a key publication on comparing randomised and non-randomised studies (Benson & Hartz 2000). A limitation of the search to these sources is deliberate, given the intention to select studies for later comparisons with randomised controlled trials and the inevitably low positive predictive value of the search strategy. Again, given the exploratory nature of this work, learning from potentially high-quality examples of ESDs was of more interest than a comprehensive retrieval of all existing published and unpublished ESDs. A restriction to English-language articles was necessary because of my own resource constraints.

**Table 6.1: Main search strategy for effectiveness studies using secondary data**

| Search | | Hits |
|---|---|---|
| #1 | Search "CHEMICALS AND DRUGS CATEGORY" [MESH] AND (EFFECTIVENESS OR EFFICACY OR OUTCOME OR EFFECT OR EFFECTS) AND (DATABASE OR DATABASES OR DATABASES OR DATABASE OR REGISTER OR REGISTRY OR REGISTERS OR REGISTRIES OR DATASET OR COHORT OR "CLAIMS DATA" OR "SECONDARY ANALYSIS") Limits: English, Human, Core clinical journals | 4452 |
| #2 | Search #1 Field: All Fields, Limits: English, Randomised Controlled Trial, Human, Core clinical journals | 498 |
| #3 | Search #1 Field: All Fields, Limits: English, Meta-Analysis, Human, Core clinical journals | 198 |
| #4 | Search #1 NOT (#2 OR #3) Field: All Fields, Limits: English, Human, Core clinical journals | 3760 |

The NHS Information Centre operates a directory of clinical databases in the UK (DoCDAT), and it was hoped that effectiveness studies which had used those databases could be identified. All names of the 44 registered clinical databases were searched for on Medline, but no additional study reference was identified which met the inclusion criteria. A later personal communication (September 2003) with one of the DoCDAT researchers confirmed this finding.

Several earlier searches had been undertaken before the main search strategy (in particular, an earlier search used the subject heading "DRUG THERAPY" instead of the more appropriate "CHEMICALS AND DRUGS CATEGORY"); additional references identified from those preliminary searches as well as references already known to the researcher have been included in the review. The fact that the main strategy had missed some of these studies is an indication of its limited sensitivity.

### 6.4.2 Inclusion and exclusion criteria for ESDs

I scanned titles and abstracts of the 3,760 identified studies and applied the criteria listed below. The criteria show the intention to identify studies with similar features as the Dornase Alfa Case Study described in Chapter 3 to 5.

**Inclusion criteria:**

1. Study uses patient-level data which has been routinely collected as part of an electronic patient record system, a registry or database
2. Study evaluates effectiveness of a particular drug intervention
3. Study includes a comparison group, i.e. either comparison between different intervention groups or intervention and control groups
4. Study is published in English.

**Exclusion criteria:**

1. Data were collected for the exclusive purpose of addressing the objectives of the reported study or another single research hypothesis
2. Reviews of patient records (paper records), if they are not part of an existing electronic patient database
3. Studies including any additional data collection from e.g. paper records, except for validation of the routinely available electronic data
4. Uncertainty regarding the inclusion criteria

   If the abstract of a paper is unclear as to whether the data stem from a database, and the total sample size is <500, the paper is excluded.
5. Safety studies (adverse events as the only outcomes)
6. Studies evaluating groups of drugs (e.g. anti-hypertensives) or combination therapies (e.g. "chemotherapy") without presenting results for individual drugs
7. Dosing studies, i.e. comparisons of groups using different doses of the same drug
8. Studies comparing use of the same drug in different contexts

### 6.4.3 Data extraction

I extracted key information from included studies using a form devised in Microsoft Excel. The information items extracted are presented in Table 6.2.

**Table 6.2: Information extracted from published ESDs**

| |
|---|
| **General information about the study** |
| Author |
| Publication year |
| Full reference |
| Affiliation and contact details of first author |

**Inclusion criteria of search**
Comparison group
Patient-level data
Routinely collected data only used
Efficacy / effectiveness of particular drug evaluated
Trial data used / included?

**Information on the data source used**
Name of database
Purpose of database
Patient numbers in database

**Stated main aim of the study**

**Study design**
Cohort / case-control
Retrospective / prospective
Historic or concurrent controls
Length of follow-up
Study funding
Source of comparison group

**Ethics and data protection**
Research ethics committee approval or equivalent
Degree of data anonymisation
Patient consent
Authors' involvement with database / registry

**Internal validity**
Patients "naïve" to treatment?
Primary and other outcomes assessed
Definition of variables
Measurement of variables
Blinding of patients and assessors
Validation and quality checks reported
Patient numbers extracted for study
Patient numbers analysed
Loss to follow-up
Data processing

**External validity**
Countries participating
Number of centres involved
Source of population
Period of data collection
Inclusion and exclusion criteria of study
Method of sample selection
Indication of representativeness of national population

**Analysis**
Power calculation
Analysis methods used
Sensitivity analyses
Differences between groups at baseline and their control
Identification of confounders and their control
Missing data

**Main results**

## 6.5 Results

A total of 42 identified studies met the inclusion criteria (Table 6.3). The main search strategy (3,760 hits) identified 32 of the 42 studies (76%). Of the remaining ten studies, four were known to me, and six were found in an earlier search using the term "drug therapy" (2,621 hits; only three studies were identified by both searches). Twenty-nine (69%) of the included studies were published in 1998 or later, and none prior to 1989.

## Table 6.3: Search results: ESDs of drug therapies

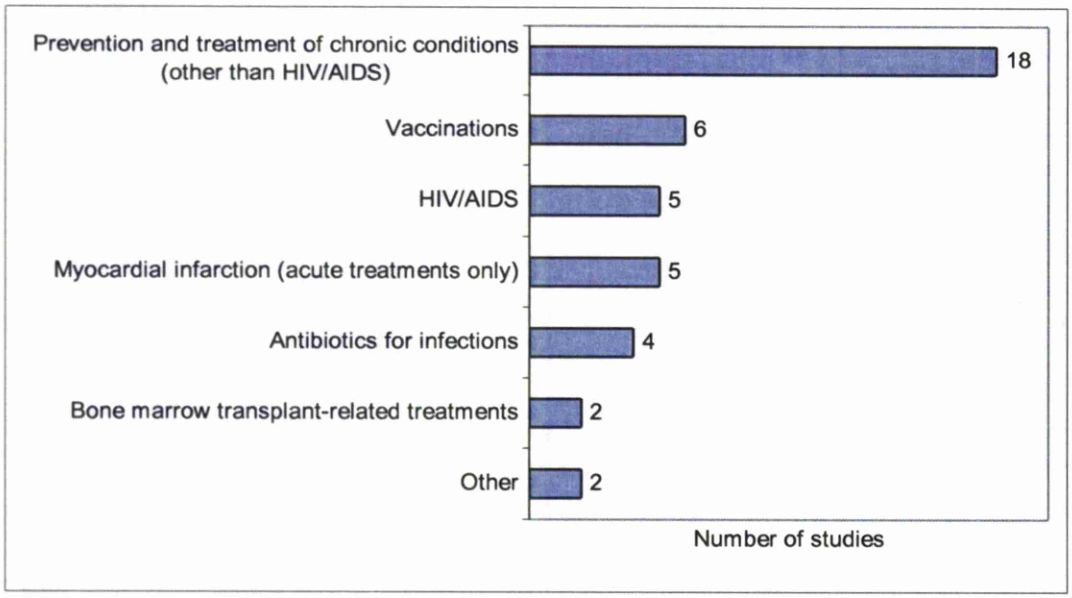
| First author | Year of publication | Assessed drug treatment(s) and condition |
|---|---|---|
| Bowman L.[2)] | 2000 | Antibiotics after treatment failure in paediatric infection |
| Benvegnu L. | 1998 | Interferon for virus-related cirrhosis |
| Berman S. | 1997 | Antibiotics in paediatric otitis-media |
| Butterworth J. | 1998 | Sufentanil, fentanyl; vecuronium, pancuronium in anaesthesia |
| Chew D. | 2001 | Clopidogrel in acute myocardial infarction (AMI) |
| Choi H. | 2002 | Methotrexate in rheumatoid arthritis (RA) |
| Eggleston A.* | 1996 | Cisapride, ranitidine, omeprazole in gastro-oesophageal reflux disease (GORD) |
| Fedson D. | 1993 | Influenza vaccination |
| Gable C. | 1990 | Pneumococcal vaccination |
| García L. [3)] | 1999 | Acid-suppressants in secondary prevention of upper gastro-intestinal bleeding (UGIB) |
| Ghani A. [2)] | 2001 | PI- and NNRTI-containing regimens in HIV |
| Giralt S. | 2000 | Interferon alfa in bone marrow transplant (BMT) |
| Goldstein R. | 1996 | Aspirin in coronary disease |
| Graham N. | 1991 | Zidovudine and (pneumocystis carinii prophylaxis (PCP) in AIDS |
| Heudebert G. [2)] | 1993 | Niacin, sequestrants, lovastatin in hypercholesterolemia |
| Huang X. [2)] | 1999 | Ciprofloxacin, nitrofurantoin, and trimethoprim/sulfamethoxazole (TMP/SMZ) for urinary tract infection (UTI) in females |
| IBMTR** | 1989 | Methotrexate in BMT for acute lymphatic leukaemia (ALL) |
| Johnson C. | 1999 | Dornase alfa in cystic fibrosis (CF) |
| Krumholz H. | 1995 | Aspirin in AMI |
| Krumholz H. | 1998 | iv heparin in AMI |
| Krumholz H. | 2001 | Aspirin (and angiotensin-converting enzyme inhibitors (ACE-I)) in AMI |
| Kuhn L. | 2000 | Zidovudine perinatally for HIV infected children |
| Lawrenson R. [3)] | 2001 | Trimethoprim, amoxicillin, cefalexin, co-trimoxazole, nitrofurantoin, cefradine, norfloxacin, ampicillin, ciprofloxacin, cefadroxil, in UTI |
| Lundgren J | 1994 | Zidovudine in AIDS |
| McDougall R. [3)] | 1994 | Prednisone in RA |
| Moore R. | 1991 | Zidovudine in AIDS |
| Nichol K. | 1994 | Influenza vaccination in elderly |
| Nichol K. | 1999a | Influenza vaccination in elderly with chronic lung disease |
| Nichol K. | 1999b | Pneumococcal vaccination in elderly with chronic lung disease |
| Nordin J. | 2001 | Influenza vaccination in elderly |
| Peterson L. [2)] | 1999 | Alteplase (rt-PA) post coronary artery bypass graft (CABG) |
| Pethica B. [1)] | 1998 | Budesonide and beclomethasone in asthma |
| Price D. [2)] | 1998 | Salmeterol in asthma |
| Rabinowitz J. | 2001 | Olanzapine or risperidone for schizophrenia |
| Rahme E. | 2002 | Naproxen to prevent AMI |
| Sebaldt R. [3)] | 1999 | Etidronate in corticosteroid-induced bone loss |
| Sernyak M. | 2001 | Clozapine on inpatient resource utilisation |
| Solomon D. | 2002 | NSAIDs in AMI prevention |
| Tiefenbrunn A. [1)] | 1998 | Alteplase in AMI |
| van Staa T. | 1998 | Cyclical etidronate in prevention of non-vertebral fractures |
| Weintraub J. | 2001 | Dental sealants |
| Ziegelstein R. [1)] | 2001 | iv magnesium in AMI |

*Paper identified published abstract of previous work which met inclusion criteria.
** International Bone Marrow Transplant Register
Identified in: [1)]main and previous search; [2)]previous search only; [3)]known to researcher (all others identified in meain search only).

### 6.5.1 Overview of conditions and interventions assessed

The majority (18) of the identified studies investigated treatments or preventative interventions (e.g. of myocardial infarction or recurrence of upper gastrointestinal bleeding) in chronic diseases or conditions such as rheumatoid arthritis, asthma, schizophrenia, cystic fibrosis, gastro-oesophageal reflux disease, or corticosteroid-induced bone loss (see Figure 6.1). A further five studies evaluated HIV/AIDS treatment. The majority of interventions were treatments, but a considerable number were of a preventative nature.

**Figure 6.1: Conditions and interventions represented in identified studies**




### 6.5.2 Data sources used by ESDs

The databases used in each study, as well as their purposes and size are listed in Table B.1 in the Appendix.

**Purpose of the databases**

The original intentions of the databases - as far as they are stated in the study reports - included the following:

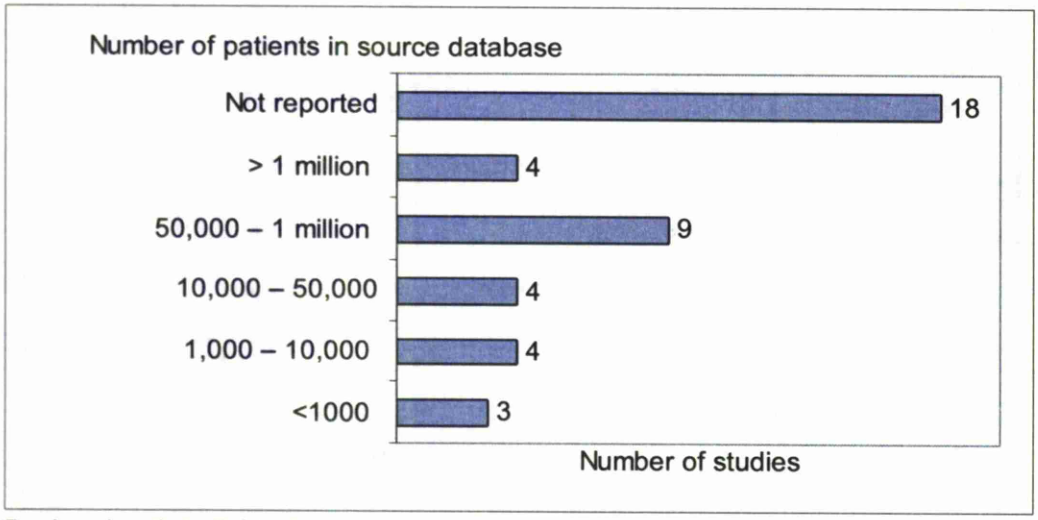
- Evaluation / description of disease process
- Description of process (including costs) and outcomes of care
- Instigation of improvements in care
- Claims data, prescription or comprehensive patient record system
- Market research
- Medical and health research (including drug safety studies)
- Evaluation of predictive potential of diagnostic tests

Fourteen studies (33%) used claims databases, of which three studies used more than one database, and two databases (GroupHealth, Inc., Medicaid) were being used by three different studies. Twenty-two studies (52%) used disease-specific data sources. Two of these databases were used three times (NRMI-2, CCP), and one twice (IBMTR). In total, 38 different data sources were used (sources of demographic statistics such as the National Death Index were not counted; different sets of Medicaid data were counted as one data source).

**Numbers of patients in database and study**

The numbers of patients in each main source database and subsequent numbers of patients sampled for the 42 studies are reported in Table B.1. Figure 6.2 provides an overview. A remarkably large proportion of studies do not describe the size of their source database (18; 43%).

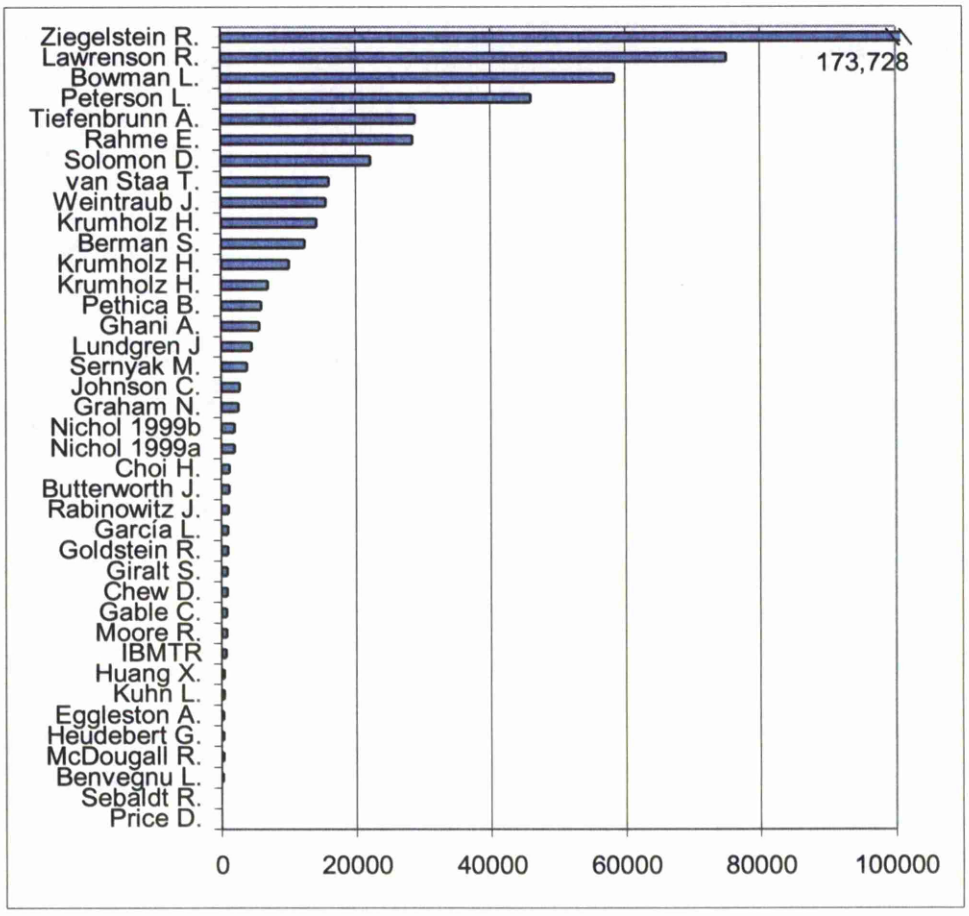
**Figure 6.2: Size of database(s) used by ESDs**




(Further details in Table B.1)

The number of patients included in the actual study analysis varied from a mere 37 (Price 1998, using a primary care database) to over 170,000 (Ziegelstein 2001) (see Table B.1). Of the four largest ESDs with known patient numbers (over 40,000 patients), two used the Second National Registry of Myocardial Infarction (NRMI-2) (Peterson, Ziegelstein), one used the General Practice Research Database (GPRD) (Lawrenson), and one used computerised prescription records of a pharmacy benefits management company (PCS HealthSystems Inc.) (Bowman).

**Figure 6.3: ESDs (n=39) ranked by reported number of included patients**




**Location of databases**

USA institutions contributed to databases used in at least 28 of the studies (for "international" studies, US participation was not automatically assumed), in 25 of these the USA were the sole contributor.

In terms of authorship, the affiliation of the first author was a USA institution in 29 studies (69%), UK-based in four studies (10%), and Canadian in three studies (7%). Of one study each, the first author was based in Belgium, Denmark, Italy, Israel, New Zealand, or Spain.

### 6.5.3 Study design

Thirty-eight studies (90%) were retrospective cohort studies, three were case-control studies with concurrent controls, and one study used both designs.

**Sample selection and selection criteria**

All but three studies based their sample selection on a set of selection criteria (Eggleston used a random sample, Goldstein and Ziegelstein used all patients in the database; the IBMTR study was not clear about whether data from all patients were analysed). Five studies selected under 2% of the database population for their study sample, seven between 2-20%, five between 20-50%; and five between 50-99%; only two studies were able to include the entire database population (details in Table B.1).

Inclusion and exclusion criteria of studies were often similar to those found in RCTs (contraindications, previous use of interventional drug, exclusion of conditions causing same or similar symptoms). However, frequent additional exclusions had to be defined because of poor or missing data or key variables of interest.

**Centres**

Twenty-five studies (60%) did not report the number of centres contributing to the data source. Five studies were single-centre studies, and for a further twelve studies, the number of centres varied between 15 and over 1,500. A large number of contributing centres, however, does not guarantee representativeness (the NRMI-2 for example is biased towards larger, academic centres). Only a minority of studies considered potential centre differences (e.g. in treatment) in the analysis.

**Comparison groups**

All studies selected their comparison groups from the same data source as the intervention group. Whereas 25 of the cohort-designs clearly used concurrent

controls, in the other 14 studies this was less clear. Many used data on patients recruited over several years, which meant that controls could potentially represent a different cohort. Only three studies adjusted for differences in calendar periods.

**Representativeness**

In only a handful of studies was it possible to conclude that the sample was representative of the national population with the condition in question (e.g. if the data source was a comprehensive national database, or representativeness had been previously examined and reported). Most studies did not comment on their likely representativeness of a particular target population, not even if the sample size and number of centres were very large. This is regrettable as it leaves the reader to assume generalisability without there being a clear case made for it.

Only very few studies gave any indication of the characteristics of excluded patients. For example, Fedson reported that they accounted for about half of influenza vaccinations dispensed in the region and assumed that most of the remainder was used for institutionalised patients which were excluded from their study. Similarly, McDougall reported characteristics of database patients excluded from the analysis. However, their database stemmed from the region's sole rheumatoid arthritis (RA) referral centre, which was estimated to see about half of all resident RA patients. This may have meant that non-registered patients were treated in primary care and might have been more mildly affected.

**Definition of variables**

The interventional variable (pharmaceutical intervention) was relatively easy to define clearly. Some studies reported treatment codes used to extract relevant records from a database. Most but not all studies required their patients to be newly introduced ("naïve") to the interventional drug or free from its use for a significant period before the observational period. In many

cases, however, this was not relevant (e.g. influenza vaccination, or acute interventions after myocardial infarction).

In terms of outcome variables, twenty-five studies (59%) defined their primary outcomes through clinical information (including hospitalisation episodes); five studies (12%) defined primary outcomes based on drug treatment data (e.g. occurrence of another treatment episode within a given time is rated as a failure of the original treatment). The remaining twelve studies (29%) used mortality or survival as primary outcome. Table B.2 lists effectiveness outcomes used in each ESD. The definitional detail varied; some studies used International Classification of Diseases (ICD) coding or required records of clinical examinations (which may have been more or less subjective).

Other variables considered in analyses were mostly only listed, rather than defined. Few studies recorded ICD codes used.

**Measurement and quality of variables**

The quality of measurements of interventions and outcomes varied also. Single-centre databases (e.g. McDougall) may have examined patients on regular follow-up visits and may even have reported the measurement instruments used, including regular calibration. In the case of multi-centre databases, the use of standardised forms for data collection was often mentioned. In claims databases, prescription records had to be extracted depending on their coding.

Surprisingly, the studies' authors rarely explicitly raised issues of data quality. Only one study reported the undertaking of consistency or range checks of data. Krumholz reported reliability testing for the interventional variable in all three studies.

Only seven studies (17%) reported having undertaken data validation against other data sources; four further studies have used data sources previously validated.

In most studies, the basis for treatment allocation (i.e. the clinicians' decision, treatment protocols, patient preferences) was completely unknown. Only one study (Choi) explicitly took account in the analysis of factors underlying the main clinician's decisions.

Patients and staff were understandably never blinded in the observational databases used, and neither were the assessors of outcomes of interest.

Most studies used variables reported in a single database. Very few studies linked data from their main data source with data from another source, e.g. for outcome measures. For example the National Death Index (USA) was used to confirm mortality outcomes (Moore, Choi), or specific hospitalisation databases were used to identify such outcomes (e.g. Rabinowitz).

**Length of follow-up**

The reviewed ESDs demonstrate little to no experience with follow-up periods which are substantially longer than in likely RCTs. Length of follow-up varied widely. The longest mean follow-up was reported by McDougall (mean 18.1 years); Weintraub followed patients for up to 8 years. Eighteen studies (43%) reported a mean, median, or overall follow-up period of two years or more. Four reported only acute treatment outcomes. Similarly, the studies evaluating influenza or pneumococcal vaccination or antibiotic treatment used comparatively short follow-up periods, typically covering some 4 months.

Some studies presented outcomes at particular follow-up intervals, with numbers of patients decreasing as the length of follow-up increased (e.g. outcomes presented for 1, 2, 3 or more years of individual follow-up);

similarly, in studies using survival analyses, a mean or median time of follow-up may have been reported; other studies presented outcomes at only a fixed time point (e.g. 1-year survival).

Differential follow-up between patient groups is a potential source of bias. In most studies, this was clearly not a problem, but in some ten studies, it was unclear whether there was such a differential. One study (Sebaldt) lost a substantial number of control patients over three years (22 of 37 patients), whereas the main outcome measure was available for all patients in the treatment group.

**Study management and funding**

The findings suggest that industry involvement in ESDs of drug therapies may be less significant than in drug RCTs. Seventeen studies (40%) had at least been partly funded by industry sources; a further three studies (7%) used the National Registry of Myocardial Infarction 2, which is industry-funded, but the authors were only identified by academic affiliations. Seventeen studies (40%) involved public funding, and seven (17%) reported financial support from charities or foundations. For six studies (14%), there was no identifiable funding source, with all authors reporting only academic or health sector affiliations.

### 6.5.4 Analysis methods used

A main criticism of ESDs is the possible confounding by indication. A variety of analytic and design control methods were represented in the reviewed studies: stratification, matching, restriction, and multivariate analysis. The majority of studies used more or less elaborate multivariate analyses to control for predictors and confounders (mostly logistic or proportional hazards regression models). Only one study used the propensity score method.

Most of those not using any control procedures were only reporting descriptive data or had a decision tree analysis as their main focus (Heudebert), or appeared to deem identified baseline differences of insufficient importance to control statistically (e.g. Huang). The control of baseline differences of course depends on the availability and quality of suitable variables, and thus the number of variables considered varied considerably between studies.

Ten studies did not explicitly report or assess differences between comparison groups at baseline; however, seven of them still used analysis methods which adjust for potential confounders. This implies a detection of baseline differences.

A third of studies took the dose or duration of the investigational treatment into account; for several others, this would not have been appropriate (e.g. vaccinations or one-off acute interventions).

Thirty-seven studies (88%) did not report any power calculation, from one study it could be inferred that such a calculation was done, and four studies explicitly reported having performed one.

### 6.5.5 Reported results

In 35 (83%) of the studies the results clearly favoured one or more of the investigational drugs over others or non-intervention control groups. These are indicated in Table 6.4 as "successful" drugs. As can also be seen on the same table, manufacturing pharmaceutical companies sponsored a minority of those studies. Twenty-nine studies (69%) did not report any study-specific involvement or funding by pharmaceutical industry, albeit that several researchers declared other financial support from such sources.

**Table 6.4: Overview of main results**

| First author | Year | Main (adjusted for confounders, if adjustment was done) result reported | "Successful" drug | Drug company involvement |
|---|---|---|---|---|
| Benvegnu L. | 1998 | Cumulative probability of worsening cirrhosis (p<0.05), development of hepatocellular carcinoma (p<0.001), and death or liver transplantation (p<0.005) was significantly lower in chronic hepatitis C patients treated with interferon | Interferon | Source of funding not reported |
| Berman S. | 1997 | Second course of antibiotics prescribed within 24 days in children with acute otitis media: mean 11.9% after Trimethoprim/ sulfamethazole (10.7%) amoxicillin (11.6%) or erythromycin/ sulfisoxazole (12.9%), and 13.2% after amoxicillin/ clavulanate (12.3%), cefaclor (13.3%), or cefixime (13.4%). p=0.029. | Cheaper antibiotics: trimethoprim/ sulfamethazole, amoxicillin, erythromycin/ sulfisoxazole | No |
| Bowman L. | 2000 | Cefaclor (referent), azithromycin (OR: 0.94, CI: 0.87-1.00), and amoxicillin/clavulanate (OR: 1.08, CI: 0.96-1.20) are associated with fewer treatment failures than other newer beta-lactams and macrolides (between 20-40%) in difficult paediatric infections | Cefaclor, azithromycin, and amoxicillin/ clavulanate | Eli Lilly |
| Butterworth J. | 1998 | Duration of intubation, ICU length of stay (LOS), and postoperative LOS did not differ between pancuronium and vecuronium; sufentanil was associated with a significant reduction (p=0.45) of 1.9 hours in duration of intubation compared to fentanyl, but no significant effect on other outcomes | Sufentanil | Source of funding not reported |
| Chew D. | 2001 | Clopidogrel is associated with substantial reduction in 30-day AMI or death risk among patients with elevated baseline C-reactive protein undergoing coronary stenting (10% vs. 24% in treated and untreated patients respectively (p=0.002), but not in the overall population 10% vs. 13% in treated and untreated patients respectively (p=0.12) | Clopidogrel | Source of funding not reported |
| Choi H. | 2002 | Mortality hazard ratio for methotrexate use vs. no use in rheumatoid arthritis: 0.4 (CI: 0.2-0.8) | Methotrexate | No study-specific funding |
| Eggleston A. | 1996 | GORD treatment prescribed for <=3 months: cisapride: 70%, ranitidine: 79%, omeprazole: 75%. Relapse rates at 6 months: 23%, 52%, and 37% respectively | Cisapride | Janssen |
| Fedson D. | 1993 | Adjusted RR of hospital admission with pneumonia and influenza following influenza vaccination: 1.50 (CI: 1.06-2.11) and 1.46 (CI: 1.06-2.11) in 1982/3 and 1985/6 respectively) | Influenza vaccination | No |
| Gable C. | 1990 | Pneumococcal vaccine efficacy varied from 50% to 75% depending on risk category | Pneumococcal vaccine | Lederle Laboratories |
| García L. | 1999 | Omeprazole maintenance therapy gives greatest protection from recurrence of upper gastrointestinal bleeding (RR: 0.2, CI: 0.02-1.0), compared to 0.9 (CI: 0.3-2.3) with cimetidine, and 0.9 (CI: 0.5-1.8) with ranitidine. | Omeprazole | Astra Hässle |
| Ghani A. | 2001 | Protease inhibitors have lower risk of failure as HIV treatment than non-nucleoside reverse transcriptase inhibitors | Indinavir, nelfinavir, ritonavir, saquinavir | Oxford Biologica |
| Giralt S. | 2000 | Short course of interferon for chronic-phase myelogenous leukemia as associated with higher risk of nonengraftment after a subsequent HLA-identical sibling bone | -- | Host of industry and other funding sources |

| First author | Year | Main (adjusted for confounders, if adjustment was done) result reported | "Successful" drug | Drug company involvement |
|---|---|---|---|---|
| | | marrow transplant (2% vs. 0.2%; p=0.01), lower risk of relapse (RR: 0.17, CI: 0.04-0.70), but does not adversely affect survival after transplant | | |
| Goldstein R. | 1996 | Cox hazards ratio for cardiac death rate in patients with coronary disease: 0.37 (p=0.023) in favour of aspirin use over 2 years | Aspirin | Ciba-Geigy, Mallinckrodt Medical, Tanabe Seiyaku Co. |
| Graham N. | 1991 | Suggests that the early primary *Pneumocystis carinii* pneumonia prophylaxis is effective in preventing first episodes of PCP in HIV-infected individuals and that zidovudine is effective in reducing the rate of progression to AIDS. | Zidovudine, PCP prophylaxis | No |
| Heudebert G. | 1993 | When lovastatin is used initially in a combination regimen for hypercholesterolaemia, the regimen becomes simpler but more expensive; at higher initial LDL-C levels the difference in cost becomes progressively smaller | Lovastatin | No |
| Huang X. | 1999 | Cure rates of uncomplicated urinary tract infection in women were 81%, 88%, and 93% respectively for ciprofloxacin, nitrofurantoin, and trimethoprim/sulfamethoxazole | Trimethoprim/ sulfamethoxazole | No |
| IBMTR | 1989 | RR of relapse after bone-marrow transplants for acute lymphoblastic leukaemia in adults: methotrexate vs. no methotrexate: 0.2 (p<0.0003) | Methotrexate | Cutter Biologicals, Sandoz, Xoma Corporation |
| Johnson C. | 1999 | Mean values of $FEV_1$ for cystic fibrosis patients treated with dornase alfa improved by 3.9% (decline in non-treated: 1.6%) - estimated benefit of 4.3% of predicted $FEV_1$ (p<.0001) | Dornase alfa | Genentec |
| Krumholz H. | 1995 | Aspirin use was significantly associated with lower mortality of elderly AMI patients (OR: 0.78, CI: 0.70-0.89) | Aspirin | No |
| Krumholz H. | 1998 | iv heparin was not associated with improved 30-day mortality rate (OR: 1.02, CI: 0.87-1.18) of elderly AMI patients | -- | No |
| Krumholz H. | 2001 | Aspirin and ACE-inhibitors significantly associated with lower 1-year mortality of elderly AMI patients (risk ratios: 0.86, CI: 0.78-0.95 and 0,85, CI: 0.77-0.93 respectively) | Aspirin and ACE inhibitors | No |
| Kuhn L. | 2000 | Zidovudine exposure was associated with higher risk of death or AIDS (OR: 1.8, CI: 1.02-3.11) in children exposed to the drug during prenatal and perinatal periods | -- | No |
| Lawrenson R. | 2001 | Young women with urinary tract infection treated with co-trimoxazole were significantly less likely to need a second course of antibiotics compared to trimethoprim (hazard ratio: 0.84 (CI: 0.75-0.95) | (Co-trimoxazole, but little difference between all investigated drugs) | Source of funding not reported |
| Lundgren J | 1994 | Zidovudine initiated after AIDS diagnosis was associated with improved survival but for no more than 2 years after initiation | Zidovudine | No |
| McDougall R. | 1994 | No long-term benefit of prednisone in rheumatoid arthritis could be demonstrated | -- | No |
| Moore R. | 1991 | Zidovudine has substantially contributed to an improved survival of AIDS patients, relative hazard: 0.62 (CI: 0.49-0.79) | Zidovudine | Burroughs Wellcome Company |
| Nichol K. | 1994 | Following influenza vaccination of elderly, the hospitalisations for pneumonia and influenza were reduced by 48-57% (p<=0.002); for respiratory conditions by 27-39% (p<=0.01); mortality decreased in vaccinated by 39-54% (p<0.001) | Influenza vaccination | Connaught Laboratories |

| First author | Year | Main (adjusted for confounders, if adjustment was done) result reported | "Successful" drug | Drug company involvement |
|---|---|---|---|---|
| Nichol K. | 1999a | Following influenza vaccination of elderly with chronic lung disease, adjusted risk ratio for hospitalisation for pneumonia and influenza was: 0.48 (CI: 0.28-0.82) p=.008; hospitalisation for all respiratory conditions: 0.76 (CI: 0.53-1.09) p=0.13 | Influenza vaccination | No |
| Nichol K. | 1999b | Following pneumococcal vaccination of elderly with chronic lung disease adjusted risk ratios of hospitalisations for pneumonia and influenza over 2 years: 0.57 (CI: 0.38-0.84); p=0.005 | Pneumococcal vaccination | No |
| Nordin J. | 2001 | Hospitalisations for pneumonia and influenza prevented following influenza vaccination of elderly: 19-20% in 1996/7 and 18-24% in 1997/8 | Influenza vaccination | No |
| Peterson L. | 1999 | No significant difference in in-hospital mortality rate or the combined end point of death and nonfatal stroke between rt-PA and PTCA in patients with acute myocardial infarction and prior coronary artery bypass graft surgery | -- | No study-specific funding |
| Pethica B. | 1998 | Budesonide has about 2/3 of the potency of beclomethasone in asthma | Beclomethasone | No study-specific funding, but other drug company support |
| Price D. | 1998 | Switch to salmeterol is cost-effective for primary care management of asthmatics | Salmeterol | Glaxo Wellcome |
| Rabinowitz J. | 2001 | No difference in re-hospitalisation rates of schizophrenic patients following risperidone or olanzapine, but lower rates than conventional antipsychotics | Risperidone, olanzapine | Source of funding not reported |
| Rahme E. | 2002 | Compared with other NSAIDs, concurrent exposure to naproxen has a protective effect against AMI on older people | Naproxen | No study-specific funding, but other drug company support |
| Sebaldt R. | 1999 | Etidronate increased lumbar spine bone mineral density (by 5.2% (p=0.016) over 3 years) in patients with established corticosteroid induced osteoporosis; significant advantage over non-etidronate users | Etidronate | Proctor and Gamble |
| Sernyak M. | 2001 | Schizophrenia patients taking clozapine recorded 33 days (36%) fewer inpatient days (p=0.0002) over 3 years than comparison group | Clozapine | No |
| Solomon D. | 2002 | No protective effect against AMI of NSAIDs generally was found; naproxen appears to be associated with reduced rate of AMI (OR: 0.84, CI: 0.72-0.98, p=0.03) | Naproxen | No |
| Tiefenbrunn A. | 1998 | Suggests that in non-shock patients eligible for thrombolysis, rt-PA and PTCA are comparable in terms of in-hospital mortality, combined end point of death and non-fatal stroke, and re-infarction | -- | No study-specific funding, but other drug company support |
| van Staa T. | 1998 | Osteoporosis patients taking etidronate had a significantly reduced incidence of non-vertebral fractures and hip fractures relative to the control group | Etidronate | Proctor and Gamble |
| Weintraub J. | 2001 | Sealants were effective in preventing carries-related services involving the occlusal surface of permanent first molars | Sealants | No |
| Ziegelstein R. | 2001 | Magnesium in AMI was associated with increased mortality (OR: 1.25, CI: 1.12-1.34) | -- | No study-specific funding, but other drug company support |

### 6.5.6 Ethics and data protection

Ten studies (24%) clearly had been approved by a research ethics committee or equivalent (studies using the General Practice Research Database (GPRD) were automatically included in this number, because they were known to have needed approval by a specific ethics committee). The remaining studies reported no ethics committee approvals.

Published reports did not clarify whether anonymous data (i.e. where no decoding key existed anywhere) had been used, but in most studies this was would hardly have been possible, given their longitudinal design. Moreover, authors of several studies included clinicians of the centres where databases or registries were compiled, and it must be assumed that at least some authors were able to identify included patients.

Most studies did not comment on data protection, anonymity, or confidentiality, or any attempts to protect them.

Only one study explicitly reported to have sought patient consent, another study sought consent for patients registered on the database, and a third study reported to have been exempt from the requirement of consent by a "university human subjects committee".

## 6.6 Discussion

Notwithstanding issues of possible publication bias, the identified 42 ESDs are relatively recent (1989 onwards) and could be seen to indicate an increase in publications of this nature. The increasing computerisation of databases and clinical systems must be expected to speed this up further. Both clinical databases and claims databases were equally represented as data sources for ESDs. A small number of databases were used for several published ESDs (e.g. GPRD, NRMI-2, Medicaid data). There is a strong USA-predominance in the running of, contributing to, and analysing of identified databases.

It is interesting to note the findings of a review of the potential use of routine databases in UK health technology assessments, which was published after this work was completed (Raftery *et al.* 2005). The review (which was not limited to drug effectiveness studies) also found very few published effectiveness studies having used UK-based routine databases. It furthermore noted that drug treatments so far were the bulk of treatments amenable to health technology assessment using UK routine databases.

ESDs vary considerably in their size (from 37 to over 170,000 patients). Of the 32 ESDs with known patient numbers, 22 (69%) included more than 1,000 patients, and only 10 (31%) over 10,000 patients. This was to a large extent a reflection of the size of the databases used. It shows that ESDs are not as a rule much larger than RCTs. In addition, plenty of smaller ESDs seem to be published in high quality journals.

The majority of studies were retrospective cohort studies and were concerned with preventive or treatment interventions for chronic conditions, which by their nature make long-term evaluations desirable. The follow-up periods ranged from several days (following acute

interventions) to an exceptional and not unproblematic 18 years in a single-centre study. However, there was next to no valuable experience with truly long follow-up periods and how these might be usefully managed in an analysis to take account of changes over time, such as involved in cohort and period effects.

The external validity of ESDs was rarely discussed or demonstrated by their authors. It may suffer because of poor availability and completeness of data needed for meaningful analysis, which necessitates pragmatic exclusion criteria. However, current ESDs could exploit their strengths more by explicitly addressing questions of external validity and generalisability.

The level of detail of any description of data collection and processing procedures varied between studies. Authors seemed to focus on potential sources of error and thus may mention the use of trained data abstractors or standard data report forms, or even a description of measurement processes and instruments. However, further discussions of data quality were sparse, and only a minority of studies used data which had been assessed for its validity or reliability. This is a serious concern, considering that our Dornase Alfa Case Study identified a plethora of data quality issues, although we dealt with a database which was routinely subjected to data checking processes!

Potential sources of bias were less frequently addressed. This may be because studies drew all patients from the same source database, and thus data definitions and measurements would have been assumed to be uniform for all patients. However, other sources of bias would have remained, e.g. where outcome assessors were aware of the study objectives, or due to differential detection or loss to follow-up.

Adjusting for confounders, particularly confounding by indication, was a prime concern of analysts in reviewed ESDs, and most presented an

assessment of baseline characteristics of the analysed patient groups. The adjustment for predictors and confounders was primarily undertaken by statistical methods and less frequently by study design features such as matching. However, the success of these adjustments was very rarely assessed or reported. Treatment allocation decisions or centre differences in practice were very rarely considered. Again, the Dornase Alfa Case Study showed that centre differences would have been an absolutely vital consideration in any multi-variate analysis due to differences in interpretation of definitions, clinical practice, and reporting practice.

There is an indication that industry funding may be less prevalent in drug ESDs compared to RCTs; also manufacturing pharmaceutical companies sponsored a minority of the 35 studies favouring a particular drug. It is of course possible that non-company studies favour non-RCT methods for financial reasons. However, post-marketing surveillance databases are frequently industry-funded and represent a significant potential resource for ESDs in future. This and the vulnerability of ESDs to selection and other biases should underline the importance of more rigorous and objective quality control of published ESDs.

### 6.6.1 Methods and limitations

The positive predictive value of the main search strategy was only 0.85% (=number of records identified by the search which met the criteria divided by all records identified by the search). Systematic searching for ESDs is at least as difficult as for observational studies, not least because there is a lack of relevant search terms in bibliographic databases. It is thus almost inevitable that relevant studies will have been missed.

Nevertheless, it was surprising that the separate search for studies using databases registered in the Directory of Clinical Databases (DoCDAT) retrieved not one study. This gives rise to the hypothesis that few such

studies have so far been conducted on UK clinical databases (a notion later confirmed by Raftery *et al.* 2005).

The restriction of the search to Abridged Index Medicus journals - although defensible - may mean that the identified studies were not representative of other ESDs published elsewhere, or indeed unpublished studies. ESDs may find it more difficult to achieve publication in a top-ranking journal. Any review should consider the potential for publication bias, as studies with significant results are more likely to be published and more likely to be published faster than those without significant results (Sterne *et al.* 2001). Particularly in the area of drug effectiveness research, the commercial interests of commissioning pharmaceutical companies contribute substantially to the decision on whether to publish any results at all (Sterne *et al.* 2001). Whereas there are statistical and graphical methods to explore the likelihood of publication bias for systematic reviews, this review was limited to reporting on issues which might be considered relevant, such as study size, significance of the results, and funding organisation.

In 35 (83%) of the studies the results favoured one or more of the investigational therapies over others or non-intervention control groups. This high proportion may indicate that a degree of publication bias was indeed at play and that ESDs with non-significant results were less likely to be published. This would not be surprising, as the same is true for RCTs (Sterne *et al.* 2001). However, selective reporting of outcome measures may still give rise to additional reporting bias. Large databases permit a wealth of analytical approaches, and it is impossible for readers of ESD reports to estimate that bias, unless they know the capabilities and parameters of the data source.

Systematic reviews usually involve a team of researchers, and many steps in the review process are double-checked or undertaken by two researchers independently to improve the validity of the process and outcomes. This

review was undertaken by only myself and may thus be more prone to error.

It was realised from the outset that this review could not be systematic, and existing resource constraints dictated pragmatic choices. For example, the exclusion of papers whose abstract was unclear as to whether the data stemmed from a database, and whose total sample size was <500 (exclusion criterion 4), constituted a pragmatic step which was a practical necessity given the number of observational studies which did not clearly identify the nature or sources of their data. Not infrequently, even the full paper did not permit a clear decision as to whether all or any data were collected specifically for the published study, or whether all data had been collected prior to the conception of the study.

The review was limited to drug interventions, and studies based on reviews of paper records were excluded. These restrictions, which have been applied to parallel the parameters of the Dornase Alfa Case Study, mean that the findings may not be applicable to studies of other interventions (e.g. surgical) or using other methods.

### 6.6.2 Concluding remarks

This review has described some of the key features of existing published drug ESDs. The findings do not encourage confidence in the quality of such studies and indicate in any case that their reporting is generally poor and usually does not equip the reader to assess their quality or representativeness fully. Many reports did not concern themselves with some of the key lessons I had identified in the Dornase Alfa Case Study (data access, selection of variables and patients, data management and quality, or data protection). Disappointingly, these studies added very little in terms of innovative solutions.

The work confirmed - and to some extent also quantified - some of the criticisms and strengths of database effectiveness studies referred to in the methodological literature. ESDs share many potential biases with other study designs, including RCTs (e.g. attrition bias, detection bias). The main concern, confounding by indication, is commonly controlled through statistical adjustment and sensitivity analyses. However, there may be doubt over the extent to which selection bias has been controlled. Furthermore, authors left much relevant methodological detail unreported. Therefore study evaluations, including an assessment of their generalisability, remain difficult, and potential strengths of ESDs, such as greater external validity, are often not fully exploited. If existing recommended reporting formats were followed (Huston & Naylor 1996, van Elm 2007), this would go some way towards increasing the value of such publications.

A carefully conducted and reported ESD should be able to make a valuable contribution to the body of effectiveness evidence around a particular treatment. The next chapter assesses the comparability of ESDs with RCTs on a sample of four case studies, and the contribution ESDs have made to the relevant body of knowledge.

# 7 Comparisons of ESDs with RCTs

## 7.1 Introduction

A small number of the ESDs identified in the review in Chapter 6 were subjected to case study comparisons with randomised controlled trials which had addressed a comparable research question. The intention was to contribute to the number of case studies of comparisons of randomised and non-randomised studies and particularly see whether there are ESD-specific lessons.

Britton et al. (1998) have suggested previously that the outcomes of RCTs and non-randomised studies would best approximate each other when the same selection criteria were used and differences in prognostic factors were adjusted for in non-randomised studies. The case studies aimed to explore this notion for ESDs.

## 7.2 Aim

Assess whether ESDs and RCTs show comparable results when selection criteria are the same and baseline differences in ESDs have been adjusted for.

## 7.3 Objectives

1. Assess study quality;
2. Assess comparability of ESDs and available RCTs;
3. Compare effect sizes of ESDs and comparable RCTs;
4. Explore reasons for any difference in findings and direction of differences;
5. Explore the contribution of ESDs to a relevant body of effectiveness evidence.

## 7.4 Methods

### 7.4.1 Selection of case studies

ESDs from the review described in the previous chapter were selected for case study comparison if the efficacy of the assessed drug treatment had also been the subject of a published Cochrane Collaboration review. Tapping into this ready resource avoided the need to undertake a full systematic review of the trials literature for each case study, as Cochrane reviews involve comprehensive systematic searches of international trial databases, as well as grey literature and in many cases hand searching. Four such comparison case studies were identified.

### 7.4.2 Assessment of study quality

The main purpose of quality assessment of studies is to guide the interpretation of the results, particularly when comparing findings between studies of different types, and to limit bias (Cochrane Collaboration 2001). Study features to be assessed according to the Cochrane Collaboration are internal and external validity, and certain design characteristics that affect interpretation of results.

**Randomised controlled trials:**

For the quality assessment of RCTs, an existing instrument was selected (Jadad *et al.* 1996). This is one of the most widely used and validated instruments available. It is also quick to use (the authors suggest no more than 10 minutes). By means of three questions, the instrument assesses whether a study was described as randomised, double blind, and whether withdrawals and dropouts were described. Scores are adjusted depending on the appropriateness of any described methods of randomisation or blinding.

**ESDs:**

There were few rigorously developed and validated instruments to assess the quality of observational studies. Such instruments included: Downs & Black (1997) (adapted by MacLehose *et al.* (2000), who suggested a further review of the instrument), Zaza *et al.* (2000), and the Newcastle Ottawa Assessment Scales (Wells *et al.* 2003). Others, such as the York Centre for Reviews and Dissemination (CRD 2001), and Cochrane Effective Practice and Organisation of Care Review Group (EPOC 1998) have suggested items to be included in quality assessments of observational or non-randomised studies.[3] Appendix C includes an overview table of the items assessed by different instruments and checklists. Existing instruments and methodological literature on ESDs (Huston & Naylor 1996; Motheral & Fairman 1997) suggested the following elements for quality assessments of ESDs:

- Internal validity
- External validity
- Data processing
- Statistical analysis
- Reporting of the study

[3] Since this work was undertaken, the CONSORT group has developed guidance on reporting observational studies (von Elm *et al.* 2007), and Deeks *et al.* (2003) have reviewed quality assessment tools for non-randomised studies. They considered only six of 194 identified tools as suitable for systematic reviews.

As far as possible, questions from existing validated instruments were used for quality assessment. In addition, questions on issues specific to ESDs using secondary data were formulated. Some of these issues have been raised in the literature (Huston & Naylor 1996; Motheral & Fairman 1997), others have arisen during the work reported here.[4] The Schedule was revised after having been piloted on two non-selected ESDs. No overall scoring was intended.

### 7.4.3 Data extraction

A data extraction sheet was developed, based on the types of information suggested for extraction in published guides to systematic reviews (EPOC 1998; CRD 2001) (see Table 7.1 for an overview of items extracted). The form was piloted on four ESDs not included in the case studies and revised following the pilot. For each of the four comparison case studies, the data extraction sheet needed to be adapted to take account of the specific disease areas and relevant patient and treatment characteristics. In particular, the collection of results had to be case study specific; only results on outcomes reported in the ESDs were collected (see Appendix C for a generic version of the form which highlights changes made following piloting). The same form was used for RCTs and ESDs, with a small number of questions being relevant to only one study type (e.g. planned duration of follow-up for RCTs).

[4] see Appendix C for the *Quality Assessment Schedule* for ESDs [together with an indication of the source of each question], and the *Guide* to completing the Schedule.

**Table 7.1: Information extracted from ESDs and RCTs included in the comparison case studies**

| |
|---|
| **General information about the study** |
| Author |
| Publication year |
| Full reference |
| Affiliation and contact details of first author |
| Countries participating |
| Number of centres involved |
| |
| **Stated aim of the study** |
| |
| **Patient characteristics and setting** |
| Source population and study setting |
| Data collection period |
| Inclusion and exclusion criteria |
| Treatment and control conditions for each group |
| Sample baseline characteristics |
| Duration of follow-up |
| |
| **Intervention** |
| Dose and duration of drug treatment |
| Continuity of treatment (including compliance) |
| |
| **Outcome measures** |
| Primary and secondary outcomes |
| Assessment method |
| Assessment timing |
| |
| **Analysis** |
| Type of analysis |
| Numbers of patients enrolled and included in analysis |
| Number of patients dropped out and reasons |
| Main analysis methods |
| Variables controlled in analysis |
| |
| **Results** |
| Use of concomitant treatments in each group |
| Within and between group differences reported at different time points |

### 7.4.4 Comparison of RCTs with ESDs

Treatment characteristics, selection criteria, outcomes measured, and analysis methods used in the included studies were reviewed and compared.

The comparability of RCTs and ESDs (beyond intervention, participants, and outcome assessments) was considered in terms of selection criteria used and adjustments for baseline differences in ESDs.

To compare results between studies of different types, pooled odds ratios and confidence intervals were used. Calculations were undertaken using the statistical software StatsDirect (Version 2.4.1).

The exploration of observed differences focused on the following issues relevant to internal and external study validity:

- Inclusion / exclusion criteria used in ESDs and RCTs
- Participation of patients and centres
- Adjustment for baseline differences in ESDs
- Follow-up

Finally, an attempt was made to assess the contribution of the ESD to the relevant body of effectiveness evidence mainly by reviewing the study's stated objectives and reported findings in comparison with other evidence.

## 7.5 Results

### 7.5.1 Case studies selected

The ESDs and related Cochrane reviews selected for the four case studies are listed in Table 7.2. Two case studies related to treatments for chronic conditions (rheumatoid arthritis, and osteoporosis), one to interventions after acute myocardial infarction (AMI), and the fourth dealt with drug therapies for gastro-oesophageal reflux disease (GORD).

The identified Cochrane reviews were considered to be sufficiently recent and close to the time of publication of the relevant ESDs. In most cases, the last substantive update of the relevant Cochrane review post-dated the ESD by up to 5 years. The timely relationship between the publication of the ESDs and the relevant comparison trials varied. In two case studies, the ESD post-dated most or all relevant comparison trials (case studies 1 and 2). In case study 4, the situation was reversed: all trial publications post-dated the ESD, apart from one trial published in the same year. In case study 3, three of six RCTs post-dated the ESD by one year (which almost certainly means that the trial itself would have at least coincided with if not predated the ESD).

**Table 7.2: List of case studies: ESDs and corresponding Cochrane reviews**

| | ESD | COCHRANE REVIEW (The Cochrane Library, Issue 4, 2002) |
|---|---|---|
| **Case study 1** | **Sebaldt *et al.* (1999):** 36 month intermittent cyclical etidronate treatment in patients with established corticosteroid induced osteoporosis<br><br>**van Staa *et al.* (1998):** Use of cyclical etidronate and prevention of non-vertebral fractures | **Homik *et al.*(1998):** Bisphosphonates for steroid induced osteoporosis |
| **Case study 2** | **Tiefenbrunn *et al.* (1998):** Clinical experience with primary percutaneous transluminal coronary angioplasty compared with alteplase (recombinant tissue-type plasminogen activator) in patients with acute myocardial infarction | **Cucherat *et al.*(1999):** Primary angioplasty versus intravenous thrombolysis for acute myocardial infarction |
| **Case study 3** | **McDougall *et al.* (1994):** Outcome in patients with rheumatoid arthritis receiving prednisone compared to matched controls | **Criswell *et al.*(1998):** Moderate-term, low-dose corticosteroids for rheumatoid arthritis |
| **Case study 4** | **Eggleston *et al.* (1996):** Outcome research in gastro-oesophageal reflux disease: retrospective analysis of prescription data (Mediplus® UK) on cisapride, ranitidine and omeprazole | **van Pinxteren *et al.*(2001):** Short-term treatment with proton pump inhibitors, H2-receptor antagonists and prokinetics for gastro-oesophageal reflux disease-like symptoms and endoscopy negative reflux disease. |

## 7.5.2 Methodological quality

**a) ESDs:**

The results of the quality assessment of the ESDs are presented in Appendix C. The four observational studies assessed (McDougall *et al.* 1994; Tiefenbrunn *et al.* 1998; van Staa *et al.* 1998; Sebaldt *et al.* 1999) offered no reassurance on the validity or reliability of their intervention variables (Eggleston's (1996) study was only available in abstract form and could therefore not be subjected to detailed quality assessment.). In van Staa's

(1998) study it is conceivable that routine data cross-checking in practices contributing to the data source might have covered the variable of interest (etidronate prescriptions). The main outcome measure in this study was clearly validated in a previous study, not so in the other three assessed studies.

Only in one study (McDougall *et al.* 1994) was it clear that patients have not been exposed to the intervention prior to the treatment period in question, even though this may be assumed in most other studies.

All four assessed studies presented baseline comparisons between the observed groups. All studies used only variables measured in the source database for the analysis.

In most studies, the representativeness of the sample and even of the source database from which it was drawn was not addressed.

**b) RCTs:**

Table 7.3 presents the quality rating of RCTs, showing a wide spread of quality scores in three case studies, and more consistent quality scores in case study 2.

## Table 7.3: Quality of RCTs in case studies

| Case study reference No. | STUDY | JADAD SCORE (Jadad *et al.* 1996) |
|---|---|---|
| Case study 1: Etidronate for established bone loss | (Pitt *et al.* 1998) | 4 |
| | (Worth *et al.* 1994) | 2 |
| | (Skingle & Crisp 1994) | 1 |
| Case study 2: Alteplase versus PTCA for MI | (GUSTO IIb 1997) | 3 |
| | (Grines *et al.* 1993) | 3 |
| | (Ribichini *et al.* 1998) | 3 |
| | (Garcia *et al.* 1999) | 1 |
| Case study 3: Prednisolone for rheumatoid arthritis | (Kirwan 1995) | 5 |
| | (Stenberg *et al.* 1992) | 4 |
| | (van Gestel *et al.* 1995) | 4 |
| | (Harris *et al.* 1983) | 3 |
| | (van Schaardenburg *et al.* 1995) | 3 |
| | (Million *et al.* 1984) | 1 |
| Case study 4: Cimetidine, cisapride, ranitidine for GORD | (Hatlebakk *et al.* 1999) | 4 |
| | (Venables *et al.* 1997) | 4 |
| | (Bate *et al.* 1996) | 3 |
| | (Carlsson *et al.* 1998) | 3 |
| | (Galmiche *et al.* 1997) | 3 |
| | (Richter *et al.* 2000) | 3 |
| | (Bate *et al.* 1997) | 2 |
| | (Hallerbäck *et al.* 1998) | 1 |

### 7.5.3 CASE STUDY 1: Etidronate for established bone loss

For this case study, there were two ESDs, both having assessed outcomes following treatment of osteoporosis patients with etidronate (van Staa *et al.* 1998; Sebaldt *et al.* 1999).

The Cochrane review by Homik *et al.* (1998) included 13 randomised and quasi-randomised trials, ten of which assessed etidronate as an intervention. Of these ten studies, six assessed etidronate as primary prevention and were thus excluded from the case study. Of the four remaining trials, one was a controlled clinical trial (Struys *et al.* 1995) and thus also excluded. Thus three RCTs were included in the case study comparison (Skingle & Crisp 1994; Worth *et al.* 1994; Pitt *et al.* 1998).

**Table 7.4: Criteria for inclusion of studies in Cochrane review by Homik *et al.* (1998)**

| |
|---|
| **Types of studies**<br>All controlled clinical trials were selected for further assessment. |
| **Types of participants**<br>Studies where participants were men and/or women over the age of 18, with underlying inflammatory disorders, initiating treatment or currently being treated with systemic corticosteroids (primary or secondary prevention), and who had not received bisphosphonates in the six months prior to the start of the study. Only those trials where the mean corticosteroid dose was 7.5 mg/day or higher were used. |
| **Types of intervention**<br>Controlled clinical trials that included any of the first or second-generation bisphosphonates, alone or in combination with calcium and/or vitamin D, with the control group taking placebo, alone or in combination with calcium and/or vitamin D were included. |
| **Types of outcome measures**<br>Primary outcome: percent change in bone mineral density (BMD) at one year at the lumbar spine or femoral neck. |

#### Treatment characteristics

The duration of treatment in the included studies ranged from technically just 3 months (the minimum in van Staa's ESD) to three years (Sebaldt)

(Table 7.5). Only in Worth's paper was it clear that compliance had been assessed during the trial.

All studies except van Staa *et al.* (1998) included calcium administration for the control group. That study, however, reported that around a quarter of patients in both groups have had calcium prescribed in the year prior to baseline. Worth *et al.* (1994) included vitamin D for the treatment group, but not the control group, whereas Pitt *et al.* (1998) administered vitamin D to both groups in their study.

**Table 7.5: Treatments investigated by studies in case study 1**

| Study | Intervention (daily dose) | Control conditions | Compliance assessed | Duration of treatment |
|---|---|---|---|---|
| SEBALDT 1999 | Etidronate 400 mg for 14 days, followed by 76 days calcium carbonate (500mg) | Calcium carbonate (500-1000 mg) | Not reported | 3 years (12 cycles) |
| VAN STAA 1998 | At least one prescription for cyclical etidronate (14 days etidronate followed by 76 days calcium) | Matched controls without etidronate prescription | No | Min. 1 cycle, (mean 1.3 years follow-up) |
| PITT 1998 | Etidronate 400 mg for 14 days; followed by 97 mg elemental Ca + 400 units vitamin D for 76 days | Placebo for 14 days; 97 mg elemental calcium + 400 units vitamin D for 76 days | Unclear | 2 years (8 cycles) |
| SKINGLE 1994 | Etidronate 400 mg + calcium 1000mg daily for 14 days followed by 13 weeks of calcium 1000 mg | 1 g calcium | Unclear | 2 years |
| WORTH 1994 | Vitamin D 1000 units, etidronate 7.5 mg/kg, calcium 1000 mg | 1 g calcium | Yes | 6 months |

### Sample selection

None of the trials reported from which country patients were recruited or how many centres participated. For both ESDs, this information was available. These two studies, however, differed considerably; Sebaldt's paper was based on an analysis of 61 patients from one specialist centre, whereas van Staa's analysis of the GPRD included nearly 16,000 patients. Female patients dominated all samples (see Table 7.9).

**Table 7.6: Characteristics of included studies in case study 1**

| Study | Countries participating | Number of centres | Number of patients screened | Number of patients included | Setting |
|---|---|---|---|---|---|
| SEBALDT 1999 | Canada | 1 | Not reported | 61 | Specialist care |
| VAN STAA 1998 | UK | 550 | Not reported | 15,954 | Primary care |
| PITT 1998 | Not reported | Not reported | 136 | 49 | Not reported |
| SKINGLE 1994 | Not reported | Not reported | Not reported | 55 | Not reported |
| WORTH 1994 | Not reported | Not reported | Not reported | 40 | Not reported |

Only part of van Staa's patients were taking corticosteroids in the year prior to study entry. All other studies included only patients on corticosteroids for significant periods of time (Table 7.7). In addition, Pitt specified a certain degree of bone loss as entry criterion.

**Table 7.7: Diagnoses and previous corticosteroid treatment of patients in case study 1**

| Study | Corticosteroid treatment | Diagnosis |
|---|---|---|
| SEBALDT 1999 | Mean >10 years prior to study | Corticosteroid-induced osteoporosis |
| VAN STAA 1998 | 26% of treatment group and 14% of control group on corticosteroids during year prior to study | Authors state that in the UK the indication for etidronate is established vertebral osteoporosis |
| PITT 1998 | Mean >8 years prior to study | Asthma, polymyalgia rheumatica, systemic lupus erythematosus etc. |
| SKINGLE 1994 | Minimum 5 mg prednisolone equivalent daily (entry criterion) | Polymyalgia rheumatica, temporal arteritis, chronic obstructive airways disease |
| WORTH 1994 | Mean ca. 7 years prior to study | Asthma |

**Table 7.8: Inclusion and exclusion criteria (case study 1)**

| Study | Inclusion criteria | Exclusion criteria |
|---|---|---|
| SEBALDT 1999 | Ambulatory, corticosteroids for ≥36 months; had lumbar spine bone mineral density measurement at baseline, at 12, 24, and 36 months | Medication other than etidronate known to affect bone metabolism within 6 months of study entry; diseases known to affect bone metabolism |
| VAN STAA 1998 | None specified | None specified |
| PITT 1998 | Corticosteroids equivalent to 5-20mg prednisolone daily for ≥6 months; age: ≥30 years<br>Lumbar spine z-score ≤1 | Pre-menopausal women unless sterilised or post-hysterectomy;<br>Generalised bone disease, including rheumatoid arthritis; inflammatory bowel disease; previous treatment with bone active agents in the therapeutic range |
| SKINGLE 1994 | ≥5mg equivalent oral prednisolone | Medication interfering with bone metabolism |
| WORTH 1994 | Clinically stable asthmatics without hypoxemia; >5mg prednisone or equivalent for >9 months; "adult" | Disorders of bone and mineral metabolism other than osteoporosis, disorders of liver or renal function; drugs other than corticosteroids affecting bone metabolism; osteoporosis treatment prior to study |

The reported baseline differences between groups in each RCT were minimal. The groups in van Staa's study were comparable on basic demographic data, but available medical history of e.g. prevalence of back pain, osteoporosis, use of hormone replacement therapy or corticosteroids, as well as fractures in the year before baseline differed considerably. Sebaldt reported longer use of corticosteroids in the etidronate group (11.1 compared to 3.8 years) and significantly lower BMD on all three bone sites measured (lumbar, femoral neck, and trochanter).

**Table 7.9: Baseline characteristics of patient groups (case study 1)**

| Study | Number of patients (number completed) | | Mean lumbar spine BMD (g/cm$^2$) | | % Female | | Mean age | |
|---|---|---|---|---|---|---|---|---|
| | *TG* | *CG* | *TG* | *CG* | *TG* | *CG* | *TG* | *CG* |
| SEBALDT 1999 | 24 (24*) | 37 (37*) | 0.89 | 1.10 | 83 | 59 | 60 | 55 |
| VAN STAA 1998 | 7977*** | 7977 | n.rep. | n.rep. | 91 | 91 | 72 | 73 |
| PITT 1998 | 26 (21**) | 23 (16**) | 0.74 | 0.76 | 62 | 61 | 59 | 59 |
| SKINGLE 1994 | n.rep. (9) | n.rep. (12) | n.rep. | n.rep. | 80 (n.s. difference) | | (n.s. difference) | |
| WORTH 1994 | 20 (14) | 20 (19) | 0.72 | 0.76 | 79 | 53 | 55 | 58 |

BMD = bone mineral density, TG=treatment group, CG=control group, n.rep.=not reported, n.s.=not significant;

* patients with lumbar BMD data at end of 3 years (femoral neck data available for only 23 and 21 patients of TG and CG respectively, trochanter data for 23 and 15 patients respectively);

** patients with lumbar BMD data at end of 2 years;

*** 1828 patients were treated for 2 or more years.

**Outcomes, analysis methods, and results**

Given van Staa's use of the GPRD as a data source, the choice of fractures as outcome measures is understandable, as BMD measurements would be far less likely to be recorded. All other studies used BMD outcomes, with some having used fractures or other markers of bone metabolism (Table 7.10).

Sebaldt and Worth reported regular calibration of the measurement instrument, and Pitt performed operator precision testing by repeat measurements on volunteers. The publication by Skingle & Crisp was a letter to The Lancet and thus lacked detail.

The definition of fractures varied significantly. Worth used x-ray radiographical measurements of vertebral height, whereby a decrease by 15% or more from a previous measurement was considered a new fracture. Van Staa used International Classification of Diseases codes, and Sebaldt relied on documentation or radiological determination. Van Staa stated a priori that the number of patients incurring fractures was the outcome chosen; Worth reported the number of atraumatic fractures, and Sebaldt reported both the number of patients and number of fractures.

Van Staa determined incidence rates (number of patients with fractures divided by patient-years of follow-up). Adjusted relative rates were estimated by Poisson regression model which included selected confounders. Cumulative survival curves, and Cox proportional hazards models were fitted. The presence of a linear trend over time was estimated for each cohort using Poisson regression; the trend difference was similarly estimated, including confounders, and - if significant - the baseline fracture history.

Sebaldt used repeated measures analysis of variance (ANOVA) and analysis of covariance (ANCOVA) to test differences in outcomes within and

between groups. A slope test examined covariate effect. The Wilcoxon rank sum test was used to test differences in corticosteriod doses between groups.

A main difference between the two ESDs was thus the approach to sample selection. Whereas Sebaldt selected a small sample of patients based on available follow-up of at least three years for the main outcomes, van Staa selected a sample by treatment allocation and based the analysis on person-time of follow-up.

Only Pitt adjusted the between-group comparison of outcomes for differences in mean daily corticosteroid use during the study, using ANCOVA. Skingle used t-tests to analyse changes in mean BMD from baseline and for comparing both groups. Worth used the Mann-Whitney U-test for their analysis.

**Table 7.10: Outcome measures used by studies in case study 1**

| Study | Outcome measures |
|---|---|
| SEBALDT 1999 | BMD lumbar spine (% change in $g/cm^2$);<br>BMD femoral neck (% change in $g/cm^2$);<br>BMD trochanter (% change in $g/cm^2$);<br>Hip fractures;<br>Vertebral fractures. |
| VAN STAA 1998 | Number of patients experiencing:<br>Fracture of hip;<br>Fracture of wrist;<br>Vertebral fracture;<br>All non-vertebral fractures. |
| PITT 1998 | BMD lumbar spine (% change in $g/cm^2$);<br>BMD femoral neck (% change in $g/cm^2$);<br>BMD trochanter (% change in $g/cm^2$);<br>Serum and urine markers of bone metabolism. |
| SKINGLE 1994 | BMD lumbar spine (% change in $g/cm^2$);<br>BMD femoral neck (% change in $g/cm^2$). |
| WORTH 1994 | BMD lumbar spine (% change in $g/cm^2$);<br>Fracture incidence of lumbar spine. |

BMD = bone mineral density

Table 7.11 presents the comparison of results for those outcomes measured in both ESDs and RCTs. No pooled estimates could be calculated, because standard deviations were only available for one RCT.

**Table 7.11: Results of studies in case study 1**

| Outcomes | Study | Results | | | |
|---|---|---|---|---|---|
| | | *Treatment group* | *Control group* | *Between-group difference* | *Years follow-up* |
| **BMD lumbar spine*** | SEBALDT 1999 | 5.2 (SD 9.8) | -1.4 (SD 9.1) | n.s. | 3 |
| | PITT 1998 | 5.1 (SE 1.0) | 1 (SE 1.5) | 4.5 (SE 1.65)% P=0.007 | 2 |
| | SKINGLE 1994 | 4.8 | -0.7 | P<0.05 | 2 |
| | WORTH 1994 | 5 | -4.3 | P<0.01 | ½ |
| **BMD femoral neck*** | SEBALDT 1999 | 0.5 (SD 11.4) | -2.6 (SD 7.3) | n.s. | 3 |
| | PITT 1998 | 2.5 (SE 1.3) | 3.6 (SE 1.7) | n.s. | 2 |
| | SKINGLE 1994 | No statistically significant changes observed after 1 and 2 years in either group; both decrease of 1% at 1 year in both groups, and increase of 3% at 2 years in treatment group. | | | |
| **BMD trochanter*** | SEBALDT 1999 | 0.6 (SD 14.6) | 0.0 (SD 6.3) | n.s. | 3 |
| | PITT 1998 | No differences observed | | | 2 |
| **Vertebral fractures** | SEBALDT 1999 | 1 fracture in 1 patient (4%) | 5 fractures in 3 patients (8%) | Not tested | 3 |
| | VAN STAA 1998 | RR: 1.26, CI: 0.95-1.67; risk in etidronate group decreased significantly over time. | | | |
| | WORTH 1994 | 0 | 6 fractures in 4 patients (21%) | n.s. | ½ |
| **Hip fractures** | SEBALDT 1999 | 2 patients | 0 patients | Not tested | 3 |
| | VAN STAA 1998 | RR: 0.66, CI: 0.51-0.85, in favour of etidronate | | | |

* mean percentage change from baseline

## Discussion

### Comparability

Both ESDs used statistical adjustments for baseline differences and confounders in their analyses. Van Staa did not specify any exclusion criteria, and it is likely that their sample contained patients which would have been excluded in other studies. A further issue possibly affecting comparability of ESDs with RCTs was the lack of blinding in the single-centre study by Sebaldt; however, a calibrated instrument was used to measure BMD.

Comparison of effect estimates

For BMD percentage changes of the lumbar spine from baseline, Sebaldt observed significant differences between both groups for up to 2 years in favour of etidronate (percentage difference 6.0% (SD 17.4), p=0.011), but not after 3 years (percentage difference 3.9% (SD 21.1), p=0.167). By comparison, the three RCTs all reported significant differences between the groups at their endpoints (after ½ or 2 years). Differences between groups of trochanter or femoral neck BMD percentage change were not significant at any time point in Sebaldt's study, and no RCT found any significant between-group differences. The control group achieved remarkably positive results in some instances (e.g. in Pitt's study).

Thus, the ESD and RCTs corresponded well in terms of their overall findings. The possibility that different parts of the skeleton may be affected differently by corticosteroids has been raised previously (Schaadt 1984 referred to by Pitt). Pitt's study (which incidentally achieved a stagnation of bone loss compared to other control groups) was the only study which included vitamin D for the control group.

The largest benefits of etidronate have been observed by Worth, whose study lasted 6 months. Pitt has also suggested that the benefit of etidronate might plateau after that time. On the other hand, Worth was the only study having used continuous rather than cyclical treatment, and having adjusted the dose for body weight. It is worth noting here that continuous therapy with etidronate is now considered harmful.

The data on fractures is difficult to interpret. As mentioned above, the determination of fractures varied considerably. Unlike other studies, van Staa followed etidronate takers (and their matched controls) until 6 months after their last prescription. Fractures were sought in medical records, based on ICD-9 categories. Van Staa observed a significant reduction in fracture risks on all observed sites, but a larger reduction for non-vertebral (RR: 0.80, CI: 0.70-0.92) compared to vertebral fractures. The authors

themselves recorded several possible limitations of their ESD, including detection bias, whereby doctors may have been more likely to recognise fractures in treated than untreated patients. In addition, only a limited number of possible confounders were available in the database and could thus be controlled for.

Inclusion / exclusion criteria

Van Staa's study did not specify any inclusion or exclusion criteria but included all patients prescribed at least one course of etidronate treatment. For control subjects, the authors selected patients from the same practice, diagnosed with osteoporosis but not treated with etidronate. Arguably, this was thus the least restrictive patient sample. No exclusion criteria were specified, but a range of variables which might have been chosen as exclusion criteria were considered as potential confounders (e.g. hormone replacement therapy). By contrast, the second ESD (Sebaldt) made inclusion primarily dependent on availability of follow-up data for long-term corticosteroid users.

All studies except van Staa sought to exclude patients previously treated with drugs affecting bone metabolism or for osteoporosis. However, van Staa might have had the opportunity to identify the first course of treatment of their patients, as long as this was already recorded on the GPRD. Problems may arise, where patients have previously tried the drug or commenced regular treatment before their practice began to report data to the GPRD, although this seems unlikely (GPRD started from 1991/2).

Participation of patients and centres

Given the inclusiveness of the study by van Staa and the representativeness of their data source (the GPRD) of the UK population, it might be argued that the generalisability of their study to all etidronate takers in the UK is exceptionally high. The indication for etidronate in the UK is established vertebral osteoporosis.

By contrast, Sebaldt included only patients from their own specialist tertiary care centre, whose data were registered at a national clinical database for osteoporosis. For the RCTs, neither the recruitment source nor the setting were known, and there was no possibility to judge their representativeness of any source population.

There was considerable variation in the reported diagnoses of patients. Whereas Worth restricted their sample to asthmatics treated with corticosteroids, the other studies were not restrictive on the indication for corticosteroid use.

Adjustment for baseline differences in ESDs

Both ESDs observed considerable baseline differences between their study groups. Sebaldt reported that they had started patients who were prescribed corticosteroids on etidronate, either because of their low BMD at initial evaluation or because of accelerated bone loss on follow-up. Thus, the difference in BMD at study baseline was to be expected. No difference in age or corticosteroid use was observed. Van Staa chose to match the study groups for age, sex, and possibly medical practice and selected a range of potential confounders for statistical adjustment, including corticosteroid use, and vertebral fracture history which differed between the groups at baseline (both higher in the etidronate group).

The nature of the dataset used by Sebaldt was far more disease-specific than the GPRD used by van Staa. Sebaldt and colleagues had relatively more control over the dataset, given that they must have known the small number of patients whose data were analysed. However, they reported adjusting for duration of corticosteroid use, but did not list other factors adjusted for in the analysis. It is likely that they would have had the majority of known variables available.

Follow-up

As mentioned above, the ESDs handled the issue of follow-up in different ways. However, whereas Sebaldt had complete data for lumbar spine BMD, data for other sites were severely depleted at the end of the three-year follow-up period (from 37 to 15 patients (-60%) in the control group). Van Staa's paper, on the other hand, unfortunately limited the reporting of the numbers followed up and the changes in relative risks over time.

Skingle & Crisp also reported that only 38% of patients were assessed after 2 years, with withdrawals at least partly caused by a drop in corticosteroid dose below the required level. Pitt had data from 81% of the treatment group and 70% of the control group available for analysis after 2 years (fewer for femoral neck BMD). Worth withdrew three patients for poor compliance and three further patients because of side effects from the treatment group of 20 patients (i.e. 30%). Only one patient was withdrawn from the control group.

Contribution to effectiveness evidence

Sebaldt explicitly aimed to "determine the effect" of etidronate in corticosteroid induced osteoporosis. Sebaldt's stated motivation was to provide long-term evidence (beyond 2 years), as an initial observed increase in BMD might only be temporary. Van Staa intended to "document" fracture rates for etidronate takers and osteoporosis patients not taking bisphosphonates in primary care.

Both study authors provided several justifications, some defending an observational study design (e.g. by referring to a possible lack of generalisability of trial evidence), others stressing the need for evidence on the direct effects of treatment on fracture risk, particularly non-vertebral fractures, which had rarely been considered.

The ESDs have thus clearly sought to reach beyond existing evidence by pursuing long-term evaluations, or assessing outcomes hitherto not

evaluated. In addition, van Staa's large-scale primary care based evaluation was able to assess fracture risk, which RCTs had not been able to address sufficiently.

Little can be said about the comparability of results, as insufficient data were available, and selection criteria, and measured outcomes varied between studies and study types. However, both types of studies favoured etidronate over control treatment and thus concurred in their conclusion. There is no indication of an overestimate of effect in the observational studies.

**Concluding comments**

One of the most interesting observations in this case study is actually the comparison between the two ESDs. They vary in their declared purpose (to demonstrate effectiveness vs. describing fracture rates), as much as their size and nature: a one-site analysis of a subject-specific database, where authors quite possibly were involved in the clinical care as well as outcome measurements, compared to an analysis of a vast primary care database without even applying explicit exclusion criteria.

These examples also highlight potential biases and strengths of both studies. On the one hand, a small single-centre study suffered from a lack of assessor blinding, as well as the potentially suspicious lack of adjustments of the analysis beyond a single factor (duration of corticosteroid use), and considerable loss to follow-up on a number of outcome variables. A single-subject database should have presented the opportunity to plan subject selection, data collection, and variable definition carefully to best support the testing of a specific hypothesis. The large GPRD-based study by comparison will have had the advantage of size, but will have had little control over errors, whether systematic or random, and will have faced potentially more pitfalls than our Dornase Alfa Case Study.

Almost paradoxically, the potential strengths (long follow-up and use of an outcome measure not normally amenable to RCTs - i.e. fractures) might have constituted some of the main potential biases in these ESDs. Future methodological developments to improve the quality of ESDs will have to concern themselves with such issues specifically.

## 7.5.4 CASE STUDY 2: Alteplase versus PTCA for AMI

The ESD for this case study (Tiefenbrunn *et al.* 1998) used the NRMI-2 database to compare primary PTCA with thrombolysis using alteplase for AMI. The outcomes assessed were in-hospital mortality, stroke, and subsequent events and procedures.

The Cochrane review on primary angioplasty versus intravenous thrombolysis for acute myocardial infarction by Cucherat *et al.* (1999) included ten randomised trials; four of them tested alteplase and are thus included in this case study (Grines *et al.* 1993; GUSTO IIb 1997; Ribichini *et al.* 1998; Garcia *et al.* 1999).

**Table 7.12: Criteria for inclusion of studies in Cochrane review by Cucherat *et al.* (1999)**

| |
|---|
| **Types of studies**<br>All randomised studies comparing primary PTCA with thrombolysis were eligible if they reported clinical outcomes within the in-hospital stay, regardless of language. Double blinding is unrealistic in this situation, the use of a placebo of angioplasty being unethical. |
| **Types of participants**<br>Patients with suspected acute myocardial infarction. |
| **Types of intervention**<br>Experimental treatment: Primary balloon angioplasty without stenting<br>Control treatment: Intravenous fibrinolytic therapy with alteplase |
| **Types of outcome measures**<br>In-hospital endpoints:<br>- death from all causes<br>- reinfarction<br>- recurrent ischaemia<br>- stroke: any stroke, haemorrhagic or non haemorrhagic<br>- severe bleeding (needing at least transfusion)<br>- combined endpoint (death and reinfarction or death, reinfarction and stroke)<br>- long-term mortality (six months or 1 year). |

### Treatment characteristics

All studies included heparin (or herudin) for both treatment arms - PTCA and thrombolysis with alteplase. In the ESD, 95% of patients or more

received heparin in both comparison groups. The Gusto IIB trial only randomised the first 89% of patients to either heparin or herudin.

**Table 7.13: Treatment investigated by studies in case study 2**

| Study | Intervention | Control conditions |
|---|---|---|
| TIEFENBRUNN 1998 | PTCA (iv. heparin in 95% of patients) | Alteplase (iv. heparin in 97% of patients) |
| GARCIA 1999 | Primary PTCA by catheter-balloon (with 10,000 I.U. heparin); stents for suboptimal balloon PTCA results or flow-limiting dissections | Alteplase starting with 15mg iv bolus, then infusion of 0.75 mg/kg over 30 min (max 50mg), and 0.5 mg/kg over 60 min (max 35 mg). Simultaneous heparin 5,000 I.U. bolus, followed by cont. perfusion |
| GUSTO IIB 1997 * | PTCA | Accelerated rt-PA: 15mg bolus, infusion 0.75mg/kg for 30min, 0.5 mg/kg for 60min (max 100 mg) |
| GRINES 1993 | 5,000-10,000 units heparin -PTCA - followed by 3-5 days iv. heparin | t-PA (Activase) 100mg (or 1.26mg/kg if <65kg body weight) for 3 hrs - followed by 3-5 days iv. heparin |
| RIBICHINI 1998 | 10,000 units heparin iv; prophylactic bolus of iv lidocain (1mg/kg); PTCA; heparin infusion for 48 hrs, ticlopidine 500mg/day for 1 month | rt-PA according to GUSTO protocol |

* First 1012 patients also randomised to heparin or herudin iv.

### Sample selection

Gusto IIB was the largest of the RCTs involving 1,138 patients of 57 centres. Grines' similarly was another multi-centre trial by the Primary Angioplasty in Myocardial Infarction Study Group (PAMI), but involved only 395 patients. The remaining two trials were smaller single-centre trials. By comparison, Tiefenbrunn's ESD was based on the NRMI-2, an observational post-marketing study sponsored by Genentech, Inc.. In 50 states of the USA, over 170,000 patients admitted to hospitals because of AMI were enrolled between June 1994 and October 1995. Of these 28,757 were included in the ESD.

**Table 7.14: Characteristics of included studies in case study 2**

| Study | Countries participating | Number of centres | Number of patients screened | Number of patients included |
|---|---|---|---|---|
| TIEFENBRUNN 1998 | USA | Not reported | 38787 treated with either PTCA or rt-PA | 28757 |
| GARCIA 1999 | Spain | 1 | Not reported | 220 |
| GUSTO IIB 1997 | USA, Spain, Belgium, Italy, Germany, Sweden, Switzerland, Australia, Canada | 57 | Not reported | 1138 |
| GRINES 1993 | USA, France | 12 | Not reported | 395 |
| RIBICHINI 1998 | Italy | 1 | Not reported | 110 |

All trials had set inclusion criteria relating to AMI diagnostic criteria. Ribichini in particular aimed to exclude small posterior AMIs. This study also excluded those aged 80 years or older and selected only patients presenting within 6 hours of symptom onset. The other trials, if at all, specified a 12-hour period for this, as did Tiefenbrunn. None apart from Ribichini excluded elderly patients. Tiefenbrunn also restricted their sample to patients with sufficient follow-up data available, in order not to miss any outcomes of interest. Only patients who had not been transferred and were eligible for thrombolysis were selected.

**Table 7.15: Inclusion and exclusion criteria (case study 2)**

| Study | Inclusion criteria | Exclusion criteria |
|---|---|---|
| TIEFENBRUNN 1998 | treated within 12 hours from onset of AMI symptoms; minimum of 48 hours in-hospital follow-up (or death) | transferred from other institution, contraindication to thrombolysis |
| GARCIA 1999 | Age: ≥18 years;<br>suspected AMI, if chest pain 30 min - 5 hours, and no response to nitrates, and ST elevation ≥0.2mV in ≥2 contiguous precordial leads. | females of childbearing age;<br>Contraindications to thrombolysis, left bundle branch block |
| GUSTO IIB 1997 | ≤12 hrs of symptom onset (chest pain for ≥20min, of ST elevation of ≥0.2mV in ≥2 contiguous leads or left bundle branch block) | Warfarin use, active bleeding, history of stroke, contraindication to heparin, renal insufficiency, systolic blood pressure >200mg Hg, diastolic >110mg Hg, childbearing potential, serum creatinine >2.0 mg/dl. |
| GRINES 1993 | ≤12 hours of onset of chest pain, ST elevation ≥1mm in ≥2 contiguous ECG leads | Inability to provide informed consent, dementia, complete left bundle branch block, cardiogenic shock, higher-than-normal risk of bleeding |
| RIBICHINI 1998 | Age: <80 years;<br>≤6 hrs of symptom onset (chest pain for > 30min, ST elevation >0.1mV on ≥2 inferior leads, and concomitant ST depression in ≥3 precordial leads, totalling >3mm) | Formal contraindications to thrombolysis or heparin, cardiogenic shock, or blood pressure <80mgHg, anticipated impossibility of percutaneous femoral vascular access. |

In terms of baseline characteristics, the ESD included more severely ill patients (expressed as percentage of patients belonging to Killip class IV) in the PTCA group than were found in the trials; however, not many trials reported that data. The ESD also included considerably more female patients and the largest proportions of patients with previous AMI and diabetes.

There were some imbalances also between groups included in particularly the smallest RCT (Ribichini); e.g. more patients with previous AMI were included in the PTCA group.

**Table 7.16: Baseline characteristics of patient groups (case study 2)**

| Study | % with previous AMI | | % diabetic | | % hypertensive | |
|---|---|---|---|---|---|---|
| | *PTCA* | *Alteplase* | *PTCA* | *Alteplase* | *PTCA* | *Alteplase* |
| TIEFENBRUNN 1998 | 18.8 | 18.0 | 17.1 | 18.4 | 41.7 | 42.4 |
| GARCIA 1999 | n.rep. | n.rep. | 12 | 17 | 32 | 39 |
| GUSTO IIB 1997 | 12.9** | 14.8** | 17.5** | 13.4** | 39.8** | 38.0** |
| GRINES 1993 | 15 | 14 | 13 | 12 | 47 | 39 |
| RIBICHINI 1998 | 18.2 | 10.9 | 16.3 | 10.9 | 40 | 45.5 |
| | **% in Killip class IV** | | **% female** | | **Mean age** | |
| | *PTCA* | *Alteplase* | *PTCA* | *Alteplase* | *PTCA* | *Alteplase* |
| TIEFENBRUNN 1998 | 4.2 | 1.3 | 27.5 | 29.2 | 60.5 | 61.1 |
| GARCIA 1999 | 2 | 4 | 16 | 20 | 63* | 60* |
| GUSTO IIB 1997 | 0.9** | 0.3** | 24.6 | 21.5 | 63.5* | 61.9* |
| GRINES 1993 | n.rep. | n.rep. | 26 | 28 | 60 | 60 |
| RIBICHINI 1998 | n.rep. | n.rep. | 18 | 15 | 63.4 | 60.2 |

* median age; ** data not available for some patients; n.rep. = not reported

**Outcomes, analysis methods, and results**

Two of the smaller trials included a particularly large array of clinical outcomes, and one of them also outcomes after 1-year follow-up. Also, Gusto IIB involved a follow-up period of 30 days, and Garcia assessed all outcomes also on 6-month follow-up. Tiefenbrunn included fewer outcome measures, and a minimum 48-hour hospital stay for follow-up.

Tiefenbrunn tested differences between the two treatment groups with chi-square or t-test depending on the nature of the variables; the effect of treatment on mortality was assessed after adjustment for potential confounders in a multiple logistic regression analysis. Garcia used the same method for adjustment.

In addition to standard chi-square and t-tests, Ribichini used Kaplan-Meier curves to analyse event-free survival (only the first occurrence of multiple events was considered). Gusto-IIB used survival analyses to study the primary endpoint. Logistic regression modelling was used to assess interactions. In addition, pre-specified subgroup analyses were performed in relation to primary and secondary endpoints. Grines undertook an intention-to-treat as well as a post hoc analysis.

**Table 7.17: Outcome measures used by studies in case study 2**

| Study | Outcomes |
|---|---|
| TIEFENBRUNN 1998 | Mortality in hospital<br>Stroke<br>Subsequent events<br>Subsequent procedures |
| GARCIA 1999 | Mortality<br>Nonfatal re-infarction (defined)<br>Post-infarction ischaemia (defined)<br>Need for revascularisation procedure after initial treatment (defined)<br>*All were also assessed on 6-month follow-up* |
| GUSTO IIB 1997 | Composite outcome of death, nonfatal reinfarction, and nonfatal disabling stroke within 30 days<br>Mortality from all causes at 30 days<br>Mortality from all causes and nonfatal reinfarction at 30 days<br>Composite end point consisting of death, reinfarction, disabling stroke, and congestive heart failure (CHF)at 30 days<br>Recurrent medically refractory ischaemia<br>Major bleeding |
| GRINES 1993 | Death<br>Reinfarction (defined)<br>Stroke<br>Bleeding<br>Pulmonary oedema<br>Intubation<br>Hypotension<br>Arrhythmias<br>Pericarditis<br>Vascular surgical repair<br>Dialysis |
| RIBICHINI 1998 | ***IN-HOSPITAL:***<br>Death<br>Reinfarction<br>Recurrence of angina at rest<br>Need for new target vessel revascularisation<br>Nonfatal stroke<br>Vascular or haemorrhagic complications needing blood transfusions<br>Severe ventricular arrhythmia<br>CHF<br>Angiographic results<br>Peak creatine kinase and time to peak<br>Length of initial hospital stay.<br>***AT 1 YEAR FOLLOW-UP:***<br>Death<br>Reinfarction<br>Recurrence of angina<br>New target vessel<br>revascularisation and<br>non-target vessel revascularisation<br>CHF with admission<br>All new admissions |

Table 7.18 compares the comparable outcome measures of included studies. For in-hospital mortality, Garcia's study is the only study which individually reports a significantly lower mortality in the PTCA group compared to the thrombolysis group. However, the pooled OR of the RCTs is 0.31 (CI: 0.13-0.69), compared to an OR of 0.96 (CI: 0.83-1.12) of the ESD.

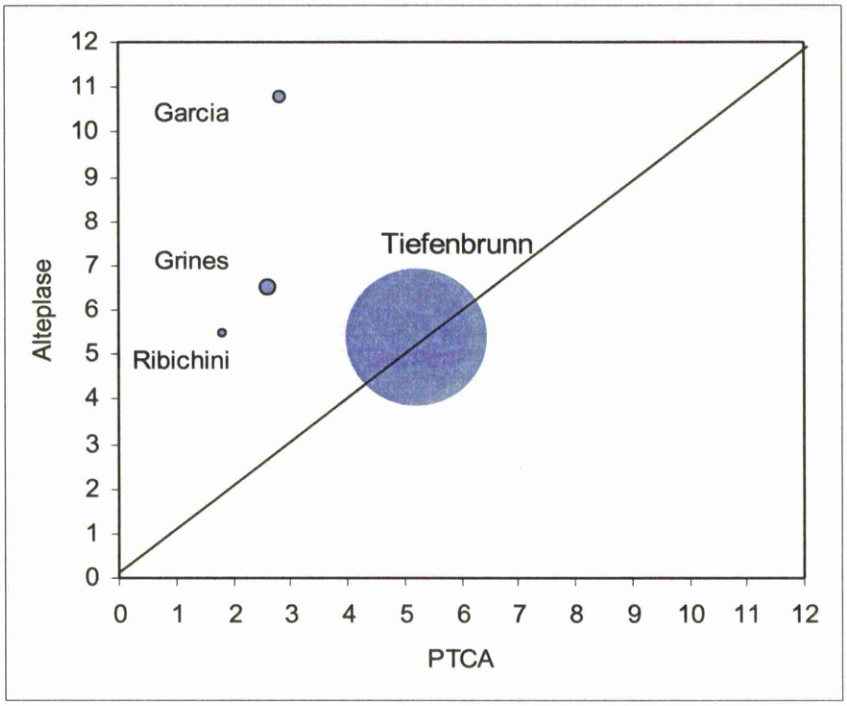
New revascularisations are defined differently in each study, and hence difficult to compare; the outcome tends to favour the PTCA group, but less clearly so in the ESD.

**Table 7.18: Results of studies in case study 2**

| Outcomes | Study | Results | | |
|---|---|---|---|---|
| | | *PTCA: Cases (%)* | *Alteplase:* Cases (%) | *Between-group difference* |
| In-hospital mortality | TIEFENBRUNN 1998# | 211 (5.2%) | 1334 (5.4%) | n.s. |
| | GARCIA 1999 | 3 (2.8%) | 12 (10.8%) | p=0.02 |
| | GRINES 1993 | 5 (2.6%) | 13 (6.5%) | p=0.06 |
| | RIBICHINI 1998 | 1 (1.8%) | 3 (5.5%) | p=0.6 |
| *30-day mortality* | GUSTO 1997 | 32 (5.7%) | 40 (7.0%) | p=0.37 |
| In-hospital stroke | TIEFENBRUNN 1998 | 173 (0.7%) | 65 (1.6%) | p<0.0001 |
| | GARCIA 1999 | 0 | 3 (2.7%) | p=0.08 |
| | GRINES 1993 | 0 | 7 (3.5%) | n.s. |
| | RIBICHINI 1998 | 0 | 0 | n.s. |
| *Stroke within 30 days* | GUSTO 1997 | 6 (1.1%) | 11 (1.9%) | Not reported |
| Reinfarction | TIEFENBRUNN 1998# | 101 (2.5%) | 716 (2.9%) | n.s. |
| | GRINES 1993 | 5 (2.6%) | 13 (6.5%) | p=0.06 |
| | GARCIA 1999 | 4 (3.7%) | 6 (5.5%) | n.s. |
| | RIBICHINI 1998 | 1 (1.8%) | 5 (9.1%) | p=0.2 |
| *Reinfarction within 30 days* | GUSTO 1997 | 25 (4.4%) | 37 (6.5%) | p=0.13 |
| New revascularisation procedure in hospital | TIEFENBRUNN 1998*# | 993 (24.5%) | 7263 (29.4%) | Not reported |
| | GARCIA 1999** | 24 (22.0%) | 53 (47.7%) | p<0.001 |
| | GRINES 1993*** | 26 (13.3%) | 126 (63%) | p<0.001 |
| | GUSTO 1997+ | 42 (7.5%) | 47 (8.3%) | Not reported |
| | RIBICHINI 1998*** | 2 (3.6%) | 14 (25%) | p=0.01 |
| | RIBICHINI 1998++ | 2 (3.6%) | 16 (29.1%) | p=0.0003 |

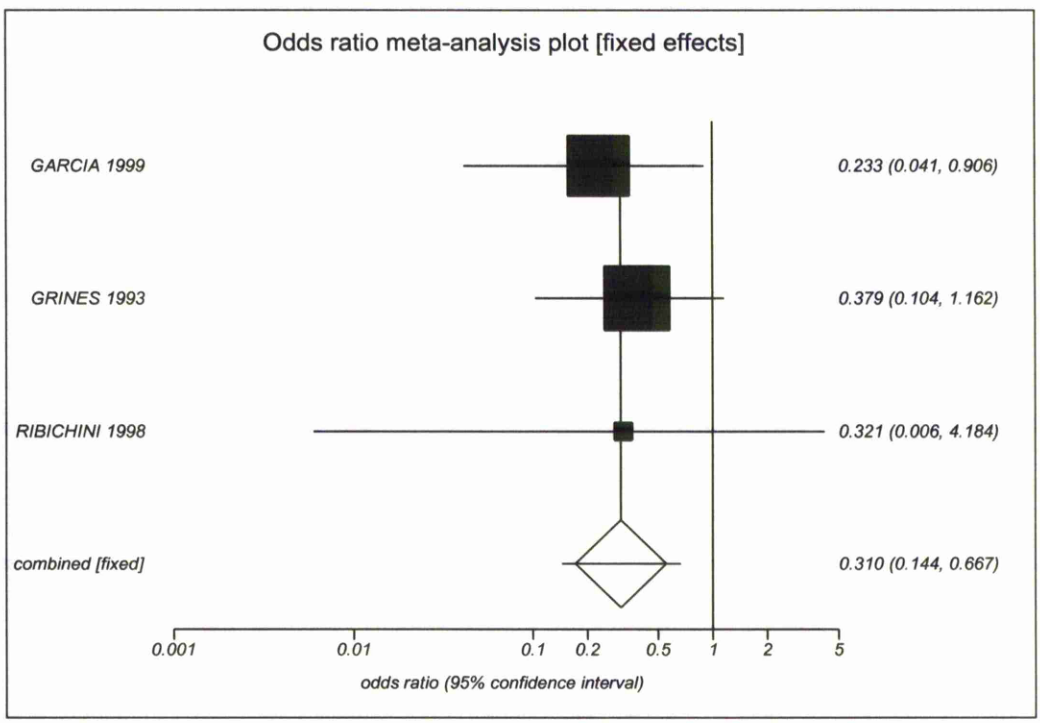
\# includes only patients not in shock
* includes: elective and rescue PTCA, elective and immediate CABG.
** includes: PTCA to left anterior descending coronary artery, or CABG
*** includes: unscheduled catheterisation only
\+ includes: CABG only
++ includes: target vessel revascularisation

**Figure 7.1: L'Abbe plot of in-hospital mortality (%) (Case study 2)**




Note: Size of points is proportional to study population.

**Figure 7.2: Odds ratio meta-analysis plot of RCTs of in-hospital mortality, favouring PTCA**




Note: ORs and CIs based on Mantel-Haenszel estimates.

**Discussion**

Comparability

Tiefenbrunn used statistical adjustments of confounding for the analysis of mortality outcome. Patients in shock were only included in the stroke outcome analysis. The exclusion of patients not remaining in hospital for at least 48 hours or having been transferred from other centres raised the possibility of selection bias. Also, whereas the authors demonstrated that transfers away fared no differently in terms of mortality outcomes, this was not assessed for patients transferred *from* elsewhere.

Comparison of effect estimates

Only one RCT (Garcia) showed a significant benefit of PTCA in terms of in-hospital mortality, but the pooled OR of three RCTs favoured PTCA. Adjusted ORs presented by Tiefenbrunn and Garcia did not alter their respective conclusions.

The difference observed by Garcia seemed to be largely due to a greater number of deaths in the alteplase group than fewer deaths in the PTCA group, by comparison to the other studies. The causes of death were mainly cardiogenic shock and free-wall rupture (5 patients each). The baseline differences of this study on a number of variables seemed to favour the PTCA group. Garcia did include 31 (14%) of patients randomised for another similar trial.

Garcia also achieved exceptionally short times from onset of symptoms to treatment: median time from onset of symptoms to first balloon inflation was 197 minutes ($25^{th}$-$75^{th}$ percentile: 150-250 minutes). Most other studies allowed up to 12 hours from symptom onset to presentation, and Garcia noted that their fast interventions resulted in the selection of relatively well patients.

Tiefenbrunn had excluded patients in shock from the main analysis. For patients in shock, mortality is significantly higher in the alteplase

compared to PTCA group (52.3% vs. 32.4% respectively: p<0.0001). This puts a different light on the comparability of the mortality results, as it is likely that two RCTs did not exclude patients in shock (including Garcia).

Also, Tiefenbrunn considered that incomplete follow-up of patients may have been a problem which affected both groups differently, as thrombolysis patients tended to be transferred earlier. All patients had a follow-up of a minimum of 48 hours (or until death), which was considered to be sufficient to capture the majority of deaths.

In-hospital strokes were rare in all studies and therefore difficult to compare. Interestingly, this was the only significant outcome in favour of PTCA reported by Tiefenbrunn in their main analysis, and the only outcome reported without excluding patients in shock. The authors suggested that the difference might have been due to having included more older patients.

Tiefenbrunn selected only those patients who actually underwent PTCA or thrombolysis with alteplase. Grines and Ribichini claimed to have performed intention-to-treat analyses and would thus have had to include patients not having received the allocated treatment (2% in Grines' PTCA group only). However, also in Gusto IIB only 82% of those allocated to PTCA underwent angioplasty. It is unclear whether they have been included in the PTCA group for analysis purposes.

Reinfarction rates did not differ significantly between the treatment groups of any study. Arguably, this was a rather less interesting in-hospital outcome, as patients could readily be subjected to further revascularisation procedures should these have been required in order to prevent reinfarction. Revascularisations were not reported uniformly, but the available data from most studies suggested that PTCA groups required significantly fewer such procedures. Again, there was a possibility that

Tiefenbrunn's study had suffered from differential loss to follow-up between the groups.

Inclusion / exclusion criteria

Tiefenbrunn noted that the mortality of patients presenting in cardiogenic shock was high in both groups, but significantly higher in the alteplase group. Two RCTs had excluded those patients. Whereas all exclusion criteria were logical, there was little consistency across studies. Garcia mentioned also that eligible patients were excluded due to physician's preference.

Tiefenbrunn had to apply pragmatic exclusion criteria (48-hour follow-up) to protect against differentially missing follow-up data. Similarly, patients transferred from other institutions were excluded.

Participation of patients and centres

The exclusion of non-thrombolysis eligible patients had significant implications for the generalisability of reported success rates from the presented studies, including the ESD. Garcia acknowledged that the profile of their sample was unrepresentative of their usual patient population. Whereas the PTCA group experienced an in-hospital mortality of 2.8%, the authors quoted a comparison figure from their clinical practice of 16.4% for patients with AMI treated with primary PTCA within 6 hours of symptom onset (and a 16% prevalence of shock at admission). They also reported that 67 patients were not randomised but referred directly to PTCA either because of exclusion criteria, patient's refusal, or physician's preference. The latter seems to imply that physicians favoured PTCA, and were willing to consider thrombolysis on equal terms with PTCA for only a selected group of patients.

In the evaluation of skill-based procedures such as PTCA, single-centre studies were rather limited in their ability to generalise their findings. The positive outcomes for PTCA in Garcia's study may not be repeatable in

centres with less experienced staff. The rate of success in performing PTCAs was similar across the RCTs (93% and more), but such rates cannot be generalised.

Tiefenbrunn briefly presented a comparison of their sample with those patients excluded because of having been transferred within 48 hours. The comparison showed some more favourable prognostic variables in the transferred patients; these patients were younger, more likely to be male or in lower Killip classes, and having experienced lower incidence of previous stroke. However, their mortality risk was similar. A preceding write-up of the NRMI reported that 14.4% of all US hospitals were represented in the registry, but these were more likely to have coronary care and invasive cardiac facilities than non-registry hospitals (Rogers *et al.* 1994).

Adjustment for baseline differences in the ESD

Tiefenbrunn used a multiple logistic regression model to adjust for potential confounders. Variables independently predictive of mortality risk were Killip class 2 or 3, age >=75 years, previous stroke, female gender, treatment interval >4 hours from symptom onset and anterior infarct location. It would appear that no such adjustment has taken place for the analysis of the other outcomes which were only reported as percentages with the level of the significance of the difference.

Follow-up

Loss to follow-up in trials was not a problem for the short-term outcomes in question. The potential for differential loss to follow-up in the ESD has been commented on above.

Contribution to effectiveness evidence

Tiefenbrunn phrased the study objective in terms of a comparison of outcomes following (or "experience" with) PTCA or thrombolysis with alteplase. However, the study was not limited to describing outcomes

following both treatments, but fully tested their equality. It is regrettable that the authors did not adhere to their original intention but rather pursued this comparative focus, which necessitated patient exclusions. They might have served the evidence base better if they had described the characteristics of patients "naturally" selected by clinicians in different centres for each treatment, and patients' experience (e.g. time to treatment, co-treatment used) and outcomes as they progressed through the treatment phases. Given the number of patients observed, this approach would have provided a rich source of naturalistic data on treatment and outcome, which could both generate clinical considerations and new hypotheses, as well as inform economic evaluations.

The literature review presented in the ESD pointed at conflicting small trial evidence, and the lack of evidence produced under non-trial conditions. Thus, a two-pronged extension to the trial evidence was aimed for: an analysis of a large number of cases, and conditions free from trial constraints. Whereas both have been achieved, neither has been fully exploited. The generalisability to a wider population of AMI patients is somewhat compromised by the fact that the NRMI-2 centres are largely specialist centres, and both interventions assessed are highly dependent on timing as well as technical skill.

**Concluding comments**

This case study is particularly interesting because of the divergent results between RCTs and ESDs. It has to be said that using conflicting small trial evidence as a justification for the Tiefenbrunn's study is somewhat unsatisfactory, as already two systematic reviews had been published at the time (Michels and Yusuf 1995, Weaver *et al.* 1997), both favouring primary angioplasty. However, the debate about the relative merits of the two alternative revascularisation methods continued for some time after that and was probably resolved in the UK in favour of PTCA at the time of the publication of the relevant HTA systematic review (Hartwell *et al.* 2005).

There are several possible reasons for the divergent findings of the ESD, and some can be glimpsed from Gusto's presentations of survival curves and sub-group analyses. These clearly show the survival advantage of PTCA starting after approximately one week (depending on outcome). Tiefenbrunn's short follow-up period could have missed much of that benefit. Similarly, there is an indication that higher-risk patients do better after PTCA; Tiefenbrunn excluded patients who were in shock (which some RCTs did too), but also those transferred from other centres. Given that the database draws on specialist units, it is possible that this would have excluded more severely affected patients, but also those with relatively late presentation (again, patients presenting relatively late have now been shown to benefit more from PTCA).

This case study thus demonstrated that selection criteria which may have been chosen for practical and unprejudiced reasons might distort the conclusions. It may also be worth noting that the authors of the ESD have consulting agreements with the company which manufactures rt-PA and funds the database (Genentech Inc.).

### 7.5.5 CASE STUDY 3: Prednisone for rheumatoid arthritis

This case study included the ESD with the longest follow-up period (McDougall *et al.* 1994). The authors termed it a case-control study, but they included only patients with rheumatoid arthritis (RA) and compared outcomes (disease activity, mortality, and adverse events) between patients treated with prednisone and those not receiving that treatment. By conventional definition, studies allocating groups on the basis of the exposure, which is followed by an outcome of interest, are cohort studies (Hennekens & Buring 1987).

A Cochrane review of "moderate-term, low-dose corticosteroids for rheumatoid arthritis" included seven RCTs (Criswell *et al.* 1998). One very early trial used cortisone acetate (Empire Rheumatism Council 1955) and was therefore excluded. Million *et al.* (1984) and Kirwan *et al.* (1995) used prednisolone; all other trials used prednisone (Harris *et al.* 1983; Stenberg *et al.* 1992; van Gestel *et al.* 1995; van Schaardenburg *et al.* 1995). Stenberg *et al.* (1992) used a complicated cross-over design, including two induction periods followed by patient-controlled medication with prednisone during two treatment periods.

**Table 7.19: Criteria for inclusion of studies in Cochrane review by Criswell *et al.* (1998)**

| |
|---|
| **Types of studies**<br>Only randomised controlled or cross-over trials of 3 months or longer were included. |
| **Types of participants**<br>Patients with a diagnosis of active rheumatoid arthritis (RA). |
| **Types of intervention**<br>Studies that used prednisone (or a comparable corticosteroid preparation) at a mean dose of less than or equal to 15 mg per day; studies that utilised either placebo controls or active controls (i.e. comparative studies). |
| **Types of outcome measures**<br>At least one of the following outcome measures in a quantitative fashion: joint tenderness joint swelling, grip strength, or erythrocyte sedimentation rate. |

**Treatment characteristics**

McDougall's study was based on the analysis of a clinical database started in 1966 in the Rheumatic Disease Unit in Saskatoon (Saskatchewan, Canada). All RA patients having been prescribed prednisone at some point, but not before study enrolment, were considered for the intervention group, regardless of treatment duration or dose. Control group patients came from the same source and were matched for age, sex, disease duration, and physician global assessment. Standard examinations and physician's global assessments (which are not further defined) on enrolment and during annual visits of patients were recorded, as were the results of semi-annually mailed Stanford Health Assessment Questionnaires.

Million used prednisolone as required for the treatment group (max. 20 mg), whereas the comparison group was allocated to bed rest and splints for inflamed joints, without steroids during active disease phases. Similarly, Stenberg permitted variable but patient-controlled doses of prednisone. Two RCTs (Harris, Kirwan) used low-dose prednisone; two further RCTs (van Gestel, van Schaardenburg) used higher doses which were then tapered off. Van Schaardenburg's RCT permitted additional chloroquine (the control treatment) if prednisone was considered ineffective.

The study duration of RCTs varied from 3 months to 10 years. Compliance testing was only reported in one study (Stenberg).

**Table 7.20: Treatments investigated by studies in case study 3**

| Study | Intervention (daily dose) | Control conditions | Compliance assessed | Duration of treatment |
|---|---|---|---|---|
| McDOUGALL 1994 | Prednisone at some time during follow-up (mean 8.0 mg) | Matched controls without corticosteroid treatment | Not reported | Mean 6.9 years |
| HARRIS 1983 | Prednisone 5 mg | Placebo | Not reported | 23 weeks |
| KIRWAN 1995 | Prednisone 7.5mg | Placebo | Not reported | 2 years |
| MILLION 1984* | Prednisolone as required | Rest and splints | Not reported | 10 years |
| STENBERG 1992 | Patient controlled prednisone medication (objective: average 2.5 mg) | Placebo | Yes | 90 days |
| VAN GESTEL 1995 | Prednisone 10mg for 12 weeks; gradually tapered off thereafter | Placebo | Not reported | 18 weeks |
| VAN SCHAARDENBURG 1995 | Prednisone 15mg for 1 month, then tapered down to effective dose; if ineffective after 3 months, or discontinued - chloroquine added | Chloroquine 100mg/day after loading during first two months. If ineffective for 3 months or discontinued - change onto gold therapy | Not reported | 2 years |

* Patients whose disease could not be controlled after min. 6 months were treated with combination of both therapies.

### Sample selection

Most RCTs were small (under 100 patients in total), except for the two UK trials (Kirwan, Million). The ESD is almost twice as large as the largest RCT in patient number terms; however, it is worth considering that the number of patient years of observation is more than six times as large (256 years for Kirwan, compared to 1684 for McDougall).

It was unclear in most studies, whether patients were recruited from tertiary or other centres (except for McDougall and Harris based in specialist centres).

**Table 7.21: Characteristics of included studies in case study 3**

| Study | Countries participating | Number of centres | Number of patients screened | Number of patients included | Setting/ Recruitment source |
|---|---|---|---|---|---|
| McDOUGALL 1994 | Canada | 1 | 1191 on database | 244 | University rheumatism centre |
| HARRIS 1983 | Pennsylvania | Not reported | Not reported | 41 | Arthritis centre |
| KIRWAN 1995 | UK | 13 | 162 "invited" | 128 | Recruitment at hospitals |
| MILLION 1984 | UK | Not reported | Not reported | 103 | Rheumatic disease centre |
| STENBERG 1992 | USA | Not reported | Not reported | 22 | Not reported |
| VAN GESTEL 1995 | The Netherlands | 1 | 55 | 40 | Recruitment at outpatients |
| VAN SCHAARDENBURG 1995 | The Netherlands | 2 | Not reported | 56 | Recruitment at hospital and clinic |

All studies took care in defining the diagnosis of rheumatoid arthritis for patient selection. The disease had to be active for recruitment into the RCTs, except for Million where treatment was administered as required whenever the disease did become active. All studies enrolled adult patients, and van Schaardenburg limited recruitment to patients of 60 years or older.

**Table 7.22: Inclusion and exclusion criteria (case study 3)**

| Study | Inclusion criteria | Exclusion criteria |
|---|---|---|
| McDOUGALL 1994 | Definite or classic RA | Prednisone before initial study enrolment; incomplete data, comorbidity requiring prednisone |
| HARRIS 1983 | Age: 18-75 years; At least 3 of 5 indices positive for activity of disease (defined) | Active peptic ulcer, renal or hepatic disease, rectal bleeding, diabetes, symptomatic idiopathic osteoporosis, cataracts, cardiovascular disease, cytotoxic treatment or levamisole, intra-articular glucocorticoid injections <6 weeks prior to study |
| KIRWAN 1995 | Age: 18-69 years; RA for <2 years and currently active (defined) | None stated |
| MILLION 1984 | Definite or classical RA for 2-24 months | None stated |
| STENBERG 1992 | Age: ≥18 years; RA according to American Council of Rheumatologists (ACR) criteria, able to give consent and understand instructions for use of prednisone, active disease (defined) | Glucocorticoid use within 6 weeks preceding the study, change in any slow-acting or investigational antirheumatic drug within past 6 months, pregnancy, range of other conditions |
| VAN GESTEL 1995 | Definite or classical RA, when treatment of gold was considered; ≥3 of 5 defined disease criteria | Diseases or medication that might affect bone mass; women ≤3 yr post-menopausal or with irregular cycles. |
| VAN SCHAARDENBURG 1995 | Definite or classic RA according to ACR criteria at age ≥60, active disease (defined), unresponsive to 3 months of NSAIDs | Use of disease-modifying anti-rheumatic drugs or prednisone during preceding 3 months, use of thiazide diuretics, inadequately controlled hypertension, peptic ulcer, hepatic, or renal disease, diabetes, ophthalmologic contraindications to chloroquine, osteoporosis, diseases precluding evaluation of therapeutic effects. |

Baseline characteristics (see Table 7.23) varied considerably in terms of disease history, with McDougall reporting on the sample with the longest disease history (around 14 years), followed by Stenberg (9.6 years). Females dominated all samples, and van Schaardenburg's sample seemed particularly poorly balanced in terms of gender. Mean ages were in excess of 48 years throughout.

**Table 7.23: Baseline characteristics of patient groups (case study 3)**

| Study | Mean duration of illness (years) | | % female | | Mean age (years) | |
|---|---|---|---|---|---|---|
| | *Treatment group* | *Control group* | *Treatment group* | *Control group* | *Treatment group* | *Control group* |
| McDOUGALL 1994 | 14.1 | 13.8 | 70 | 70 | 55.9 | 56.0 |
| HARRIS 1983 | 7.1 | 6.9 | 72 | 63 | 54.9 | 53.9 |
| KIRWAN 1995 | 1.3 | 1.3 | 62 | 66 | 48.2 | 50.3 |
| MILLION 1984* | (*2-24 month*) | | 78 | 83 | 48.0 | 47.7 |
| STENBERG 1992* | 9.6 | | 61 | | 60.9 | |
| VAN GESTEL 1995 | 1.8 | 2.5 | 70 | 70 | 57** | 56** |
| VAN SCHAARDENBURG 1995 | 0.9 | 0.8 | 71 | 43 | 69 | 70 |

*Baseline data excludes deaths and dropouts
**Median age

### Outcomes, analysis methods, and results

The variety of outcomes assessed in the different studies was staggering and ranged from generic health questionnaires, generic and symptom-specific assessments of disease progression, such as joint swelling, pain, grip strength, inflammatory markers, and subjective assessments, to bone mineral density, and mortality. A vast array of scoring systems and instruments were used (see Table 7.24). Clearly, a direct quantitative comparison of results was impossible given the diverse outcome measures.

**Table 7.24: Outcome measures used by studies in case study 3**

| | Outcomes |
|---|---|
| McDOUGALL 1994 | Physician Global Assessment (scores: 0 to 100)<br>Mortality<br>Lansbury index (summarises indices of rheumatic activity)<br>Functional class<br>Stanford Health Assessment Questionnaire<br>Side effects:<br>Fractures<br>Cataracts<br>Osteonecrosis |
| HARRIS 1983 | DISEASE ACTIVITY:<br>Pain<br>Joint swelling<br>Tenderness<br>(combination of joints evaluated)<br>Global sense of well-being<br>Duration of morning stiffness<br>Grip strength<br>Time taken to walk 15.2 meters<br>Erythrocyte sedimentation rate (ESR)<br>QUALITATIVE TESTS: |

| | Outcomes |
|---|---|
| | Subjective assessment |
| | Functional class |
| | Functional capacity |
| | Radiographic examinations of hands for soft tissue swelling, joint space narrowing, progression of erosion and osteoporosis, and lumbar spine for fractures and osteoporosis |
| | Ophthalmological changes |
| KIRWAN 1995 | Progression of damage (Radiographs of hand) |
| | Development of erosions on hands without erosions at baseline (Erosive or nonerosive, and joint destruction of each finger or wrist joint scored by Larson method (0-5) - scores from 0 - 140 for sum of joint scores) |
| | Disability (Health Assessment Questionnaire) |
| | Joint inflammation (Articular index weighted for joint size) |
| | Pain over prev. 24 hours (visual analogue scale) |
| | Acute-phase responses (ESR, C-reactive protein, or plasma viscosity, depending on centre) |
| MILLION 1984 | Duration of morning stiffness |
| | Number of inflamed joints |
| | Degree of pain |
| | Synovial thickening |
| | Range of motion |
| | Subluxation |
| | Instability and ankylosis |
| | Grip strength |
| | Functional capacity |
| | ESR |
| | Blood pressure |
| | ARA criteria |
| | Rheumatoid factor |
| | White-cell count |
| | Radiographs of hands, wrists, metatarsophalangeal joints, knees, hips |
| STENBERG 1992 | Tender joint count |
| | Swollen joint count |
| | Duration of morning stiffness |
| | Time until fatigue |
| | ESR |
| | Haemoglobin |
| | Daily total pain score (diary) |
| | Number of painful joints (diary) |
| | Extent of morning stiffness (diary) |
| | Global assessment (diary) |
| | Medication used (diary) |

| | Outcomes |
|---|---|
| VAN GESTEL 1995 | Disease activity score (A composed index containing Ritchie articular index, number of swollen joints, ESR and visual analogue scale (VAS)-self-assessed general health)<br>Patient-assessment of pain (visual analogue scale)<br>Morning stiffness (in minutes)<br>Haemoglobin<br>Thrombocytes<br>Mean grip strength<br>Functional capacity (Dutch equivalent of Stanford Health Assessment Questionnaire (HAQ) (from 0 to 3))<br>Joint erosions (independently by two observers)<br>Joint space narrowing (modified Sharp method - reference given) |
| VAN SCHAARDENBURG 1995 | Patient assessment of disease in comparison with baseline (4-point scale)<br>Dutch Health Assessment Questionnaire<br>Number of swollen joints<br>Modified Ritchie Articular Index (max 69)<br>ESR<br>Radiographs (Scoring of joint erosions and space narrowing)<br>BMI<br>Degree of osteophytosis in lumbar spine<br>Physical activity<br>Bone mineral density<br>Biochemical parameters |

In terms of analytical approaches, McDougall used standard tests to compare outcomes between matched groups: t-test, chi-square test for paired data, Fisher's exact method for comparison of proportions, and Mantel-Haenszel method for mortality rates comparisons. Thus matching seems to have been the only method for adjusting for baseline differences.

Standard statistical tests were used in the RCTs (chi-square or Fisher's exact test, t-tests, and Wilcoxon rank sum test), and in the case of Stenberg's cross-over study, analysis of variance for repeated measures. Skewed distributions were log-transformed. In addition, Kirwan took account of different methods used in different centres to measure response, by standardising values. Van Schaardenburg used analysis of variance for repeated measures to be able to use available data on patients not completing the study.

Table 7.25 attempts to give a comparative overview of outcomes between the studies. No pooling of results was possible, due to the diversity of outcome measures used in different studies.

**Table 7.25: Results of studies in case study 3**

| Outcomes | Study | Results | | | | |
|---|---|---|---|---|---|---|
| | | *Treatment group* | *Control group* | *Between-group difference** | | *Years follow-up* |
| Mortality | MCDOUGALL 1994 | (52) 42% | (67) 54% | P<0.05 | | 10 |
| | VAN SCHAARDENBURG 1995 | 0 | 4 (14%) | | | 2 |
| Stanford Health Assessment Questionnaire | MCDOUGALL 1994 | Mean 1.6 | Mean 1.3 | n.s. | | 10 |
| Health Assessment Questionnaire | VAN SCHAARDENBURG 1995 | Improvement (p<0.01) | Improvement (p<0.01) | n.s. | | 2 |
| General health | VAN GESTEL 1995 | at 12 weeks (before tapering off) significantly better in treatment group | | | (+) | 0.23 |
| Global sense of well-being (percentage of baseline) | HARRIS 1983 | Mean: 32.6 | Mean: 7.8 | P<0.01 | (+) | 0.46 |
| Functional class | MCDOUGALL 1994 | No significant difference between groups until year 10 (p<0.01) | | | (-) | |
| (Score 1-5) | MILLION 1984 | 2.6 | 3.2 | n.s. | | 10 |
| (Score 1-4) | HARRIS 1983 | Not clearly reported, but difference at 24 weeks seems n.s. | | | | 0.46 |
| Lansbury Index | MCDOUGALL 1994 | 75.6 | 57.4 | p<0.01 | (-) | 10 |
| Larsen Score (measures joint destruction) | KIRWAN 1995 | 0.02±0.43 | 0.30±0.52 | p=0.004 | (+) | 2 |
| Number of inflamed joints | MILLION 1984 | N.s. difference between groups | | | | 10 |
| Disease Activity Score | VAN GESTEL 1995 | at 12 weeks (before tapering off) significantly better in treatment group | | | (+) | 0.23 |
| Radiographic scores: Joint space narrowing, erosion score, and number of affected joints | VAN SCHAARDENBURG 1995 | Significant progression | Significant progression | n.s. | | 2 |
| Fractures | MCDOUGALL 1994 | 31 patients | 19 patients | p<0.05 | (-) | 10 |
| | VAN SCHAARDENBURG 1995 | 3 patients (11%) | 1 patient (4%) | | | 2 |

* (+)= significant in favour of treatment, (-)= significant in favour of control intervention

## Discussion

### Comparability

The comparability between the ESD and RCTs in this case study is clearly not satisfactory. It was difficult to judge whether the samples in the ESD and trials were similar. Firstly, the trials used different exclusion criteria,

and it is possible that patients who would have been excluded from trials were included in the ESD. Patients with contraindications to prednisone may in theory have been over-represented in the control group. Moreover, the ESD excluded a large number of patients who had been treated with prednisone before enrolment, which may have given rise to selection bias. Even baseline data were difficult to compare, as some trials excluded deaths and dropouts from baseline data presentation. McDougall used a matched study design, but did not account for other differences in the analysis, in particular a differential use of disease modifying drugs between the treatment and control groups.

Globally speaking McDougall's results suggested that there might be no long-term benefit from the use of prednisone in RA, as for example indicated in the significantly poorer results on physician's global assessment after 10 years. McDougall followed patients from 1966, and most of the RCTs have been published in the mid-1990s. There was no directly comparable RCT in terms of length of follow-up. Million argued that their trial tested two alternative long-term regimens for synovitis, not prednisone, and many of their patients either dropped out or moved on to a joint regimen. The 2-year trials (Kirwan, van Schaardenburg) were continuous trials and thus hardly comparable. The same applied to the remaining shorter-term trials.

Comparison of effect estimates

The incompatibilities in outcome measures across studies made any attempted comparisons difficult and synthesis impossible. The multitude of measures raises the possibility for selective reporting and thus reporting bias, which is difficult to assess. Similarly, trialists may have been tempted to perform multiple testing and thus could in theory achieve and report spuriously significant results. It would seem surprising if a clinical database like the one used by McDougall was not also recording grip strength for example. However, this was not reported.

The between-group difference in mortality observed by McDougall after 10 years was statistically significant, but the authors put this down to a difference in recruitment period between the groups. Mortality results could also be gleaned from van Schaardenburg's trial, albeit that this was not reported as an outcome measure. Four control-group patients died of causes unrelated to the control treatment (chloroquine). During Million's 10-year study, 17% (9/53) and 16% (8/50) in the treatment and control groups respectively died.

The comparatively low rates are noteworthy (McDougall: 42% and 54% in treatment and control group respectively). A simple explanation for the difference could be the age of patients, but this was impossible to determine, since Million only reported baseline data for patients completing the study. The difference in duration of illness prior to study enrolment (2-24 months in Million's study, and around 14 years in McDougall's) may indicate that the studies had indeed recruited patients from essentially different age groups. But it is also possible that McDougall's sample contained more patients who were diagnosed earlier in life and that the difference in mortality was real.

As for more disease-specific indicators, McDougall observed a statistically significantly poorer Lansbury Index in the treatment group compared to the control group after 10 years (but not 5 years). Shorter RCTs indicated more positive results for the treatment group. For example, van Gestel's treatment group fared better in terms of Disease Activity Score, but also Ritchie Articular Index, than the control group after 12 weeks' treatment; however, the difference disappeared after the treatment was tapered off. The most favourable result for prednisolone in terms of joint destruction came from Kirwan's two-year study of continuous low-dose therapy. The number of inflamed joints rated in Million's study did not differ between groups after 10 years, regardless of whether patients were considered who followed the original regimen or patients who had moved onto a joint

regimen. However, deaths and dropouts were not considered in the analysis. Results by van Schaardenburg were similar.

Global assessments of health and well-being showed no advantage of prednisone over control group outcomes in the ESD or a 2-year study by van Schaardenburg. However, this latter study was limited to elderly patients, whose co-morbidity may account for a considerable background level of poor health. Again, shorter trials by van Gestel and Harris showed more positive results for prednisone; this was similar to the findings by Stenberg's cross-over study of 90-day treatment periods. Patient-rated global assessment was significantly better during prednisone periods compared to placebo periods. However, this trial also used a washout period, whereby 3 (14%) patients not responsive to a 14-day prednisone course were excluded from the trial. Also, a large placebo effect was observed, despite significant between-group differences.

During post-treatment follow-up periods, these positive results changed considerably in both van Gestel's and Harris' studies. In the latter, the treatment group showed an improvement from baseline by +32.6% (5%CI: ±16.5) after 24 weeks' treatment, and a deterioration of -23.6% (5%CI: ±20.2) from baseline only 8 weeks later. This rebound phenomenon may explain some of the lack of benefit seen in the 10-year studies which would have involved episodic rather than continuous treatment.

Inclusion / exclusion criteria

Exclusion criteria in the RCTs - as far as stated - centred around the use of glucocorticoids within specified periods prior to study entry and conditions in which prednisolone would only be used with caution (e.g. renal impairment). Thus, neither of those criteria should have restricted the eligibility of patients considerably more than the ESD did. In fact, the latter had to exclude patients with incomplete data (4% of the study cohort). Perhaps the most restrictive exclusion criterion was used by van

Gestel, who excluded all conditions and medications affecting bone mass, without having specified these further.

Inclusion criteria were relatively homogeneous between studies, except for van Schaardenburg's sampling of patients of 60 or more years of age.

Participation of patients and centres

McDougall presented a comparison between the patients included in their study and those excluded but registered on the source database. Despite a good match between the groups in terms of age, sex, age at RA onset, and some disease markers, there were several differences observed, e.g. matched subjects had significantly longer follow-up periods. They were also more likely to be treated with disease-modifying drugs. It is possible that the included and treated patients were referred earlier to the centre and thus more likely to have been started on prednisone after enrolment to the database (starting treatment before enrolment was one of the exclusion criteria).

McDougall, however, represented 50% of their region's RA patients, despite being a single-centre study. Theirs was the only referral centre in the region, which raises the question of whether the remaining not-referred patients were systematically different from those included in the database. McDougall did not further comment on this. Neither did the RCTs allow any assessment of generalisability of their findings.

Adjustment for baseline differences in ESDs

In McDougall's study, the baseline characteristics were remarkably similar between the two groups, including disease markers not matched for, such as functional class, Lansbury index, or erythrocyte sedimentation rate. However, the only presented significant difference - a greater likelihood in the treatment group of receiving disease-modifying drugs - was not adjusted for in the analysis.

Follow-up

Of the 122 matched pairs of patients included in the ESD, only 77 and 70 patients were followed-up for 10 years in the treatment and control group respectively. However, for some outcomes, data availability was even more restricted, e.g. Physician Global Assessment was only available for 55 and 39 patients respectively. This further limits the generalisability of the case study, unless missingness can be shown to have been random.

In the other 10-year study (Million) 15% of patients died and a further 19% were lost to follow-up. None of them seemed to have been included in the analysis. Of an original 53 and 50 patients in the prednisone and rest/splint group respectively, 37 and 29 completed 10 years of treatment. Only 24 (65%) and 22 (76%) of these stayed on the original programme, with the remaining patients having moved to combined treatment.

The loss to follow-up in most of the shorter RCTs was small. However, in van Schaardenburg's study, 43% (12/28) and 39% (11/28) of patients moved from prednisone and chloroquine single-treatment respectively to various combined treatment options.

Contribution to effectiveness evidence

Similarly to the ESDs in case study 1, McDougall's objective was clearly centred around assessing long-term outcomes, and the 10-year follow-up easily exceeded the duration of available trials, apart from the rather unusual trial by Million. The confidence in the findings must, however, be limited for several reasons: firstly, this was a single-centre study based in an academic rheumatic disease unit, with thus limited generalisability. The long observation period of nearly two decades may also be detrimental to both generalisability and internal validity as management and assessment practices change. The control of possible confounding by indication was only managed by a matched design, and thus differences in supplementary treatments were not accounted for.

The case of prednisolone for RA may be similar to that of dornase alfa for cystic fibrosis (except for the more detrimental side effect profile of the former): despite encouraging short-term evidence for the treatment, no RCTs have assessed the long-term benefits. The ESD assessed relatively older and much longer-suffering patients than the RCTs, and observed them during normal clinical practice, regardless of duration or continuity of treatment. The authors caution against embracing long-term therapy, but have not clearly presented the outcomes for patients on continuous long-term therapy (some 43% of patients used prednisolone for more than 5 years), compared to maybe those with intermittent treatment. This seems a lost opportunity, and it is difficult to see what clear conclusions clinicians may be able to draw from the presentation of this study.

**Concluding comments**

Clearly this case study was not a success in terms of comparability of outcomes. However, some general lessons may be drawn from it. The generalisability of a single-centre ESD is by necessity severely limited, and it may contribute little to the effectiveness evidence base (apart from a one-centre case study). The example further highlights possible limitations of very long follow-up periods of ESDs: the patient cohort may change over time and this may impact on outcomes; similarly assessors will change and hence their potential consistency may suffer; lastly, treatment decisions (and hence patient selection for treatment) may change, e.g. due to new alternatives, and therefore may render the results ultimately irrelevant in a new context.

### 7.5.6 CASE STUDY 4: Omeprazole, cisapride, ranitidine for gastro-oesophageal reflux disease (GORD)

The ESD by Eggleston *et al.* (1996) used the Mediplus® UK database, a commercially owned database used for marketing research. A random sample of patients on the study drugs was screened for patients with a first diagnosis of gastro-oesophageal reflux disease (GORD) (*n*=279). The study has only been published in abstract form. Further data were requested from the authors directly, but were said to no longer be available.

Of 17 trials included in a Cochrane review by van Pinxteren *et al.* (2001), eight could be included in the case study (see Table 7.27). The remaining nine trials used drugs or outcome measures not comparable with those in Eggleston *et al.* (1996).

**Table 7.26: Criteria for inclusion of studies in Cochrane review by van Pinxteren *et al.* (2001)**

| |
|---|
| **Types of studies**<br>Randomised controlled trials with a single- or double-blind design, in which one of the intervention types was contrasted with placebo or another intervention type.<br><br>**Types of participants**<br>Adults, both gender<br>Predominant heartburn (a retrosternal burning sensation), diagnosed as GORD or reflux-like dyspepsia<br><br>**Types of intervention**<br>Short-term treatment (one to twelve weeks) with proton pump inhibitors (esopremazole, lansoprazole, omeprazole, pantoprazole and rabeprazole), H2-receptor antagonists (cimetidine, famotidine, nizatidine and ranitidine) or prokinetics (cisapride, domperidone and metoclopramide).<br><br>**Types of outcome measures**<br>*Primary*: heartburn remission (defined as no more than one day per week with mild heartburn)<br>*Secondary*: (partial) symptom relief; quality of life |

#### Treatment characteristics

Seven of the eight RCTs included in case study 4 assessed omeprazole, two each included ranitidine, and cisapride, and one cimetidine. Apart from

Hallerbäck (1998), the double-blind period of all other RCTs lasted for at least 4 weeks. The ESD by Eggleston (1996) - which was only available in abstract form - did not describe the duration of treatment, nor the doses used; rather, outcome assessments were made dependent on the duration of the initial treatment with one of the study drugs.

**Table 7.27: Treatments investigated by studies in case study 4**

| Study | Interventions tested (daily dose)* | Compliance assessed | Duration of treatment (double-blind phase in RCTs) |
|---|---|---|---|
| EGGLESTON 1996 | CIS, RAN, OME (doses unspecified) | No | Not reported |
| BATE 1996 | OME (20mg) - Placebo | Unclear (patients are excluded due to non-compliance) | 4 weeks |
| BATE 1997 | OME (20mg) - CIM (1600mg) | Not reported | 4 weeks |
| CARLSSON 1998 | OME (20mg) - OME (10mg) - Placebo (only endoscopy-negative patients randomised to placebo) | Not reported | 4 weeks |
| GALMICHE 1997 | OME (20mg) - OME (10mg) - CIS (40mg) | Yes | 4 weeks; (+ 4 weeks for those still experiencing symptoms) |
| HALLERBÄCK 1998 | RAN (300mg) - Placebo | Yes | 2 weeks |
| HATLEBAKK 1999 | OME (20mg) - CIS (20mg) - Placebo | Not reported | 8 weeks |
| RICHTER 2000 | OME (20mg) -OME (10mg) - Placebo | Yes | 4 weeks |
| VENABLES 1997 | OME (20mg) - OME (10mg) - RAN (300mg) | Yes | 4 weeks |

* CIM=Cimetidine, CIS=Cisapride, OME=Omeprazole, RAN=Ranitidine

### Sample selection

Only two RCTs reported the number of patients screened before randomisation; the majority recruited patients in primary care. The sample size varied from 209 to 994, with the ESD being one of the smaller studies.

**Table 7.28: Characteristics of included studies in case study 4**

| Study | Countries participating | Number of centres | Number of patients screened for inclusion | Number of patients included |
|---|---|---|---|---|
| EGGLESTON 1996 | UK | Not reported | unknown | 279 |
| BATE 1996 | UK/ Ireland | 23 | unknown | 209 |
| BATE 1997 | UK | 19 | unknown | 221 |
| CARLSSON 1998 | Australia, Holland, Norway, UK | 36 | unknown | 538 |
| GALMICHE 1997 | France | multicentre study | unknown | 424 |
| HALLERBÄCK 1998 | Sweden, Finland, Norway | 21 | 441 | 427 |
| HATLEBAKK 1999 | Norway | 65 | 573 | 483 |
| RICHTER 2000 | USA | 36 | unknown | 359 |
| VENABLES 1997 | UK | 106 | unknown | 994 |

Whereas Eggleston (1996) restricted their study to patients with the *first* diagnosis of GORD, many of the RCTs specifically selected patients with a history of complaints lasting at least one or three months. Only a minority of studies used endoscopy to verify a diagnosis or exclude oesophagitis. Thus, some RCTs included patients with endoscopic abnormalities, whereas others excluded them.

## Table 7.29: Inclusion and exclusion criteria (case study 4)

| Study | Inclusion criteria | Exclusion criteria |
|---|---|---|
| EGGLESTON 1996 | Patients with a first diagnosis of GORD occurring after 1.1.1993 and treated with any of the study drugs. | Recurrent diagnosis of GORD, H2-refractory GORD, referral prior to first treatment |
| BATE 1996 | 18-80 years; heartburn as predominant symptom; endoscopically verified normal oesophageal mucosa or erythema with no erosions. | Ulcer, duodenitis, angina, ischaemic heart disease, biliary disease, pancreatitis, inflammatory bowel disease (IBD), Barrett's oesophagus, active gastro-intestinal (GI) bleeding, varices, stricture, motility disorder, surgery, significant illness, omeprazole or H2 antagonists, anti-reflux drugs, warfarin, phenytoin. |
| BATE 1997 | 18-80 years; heartburn as predominant symptom in month prior to entry; heartburn on ≤2 or 7 days before medication; normal oesoph. mucosa or oesophagitis without frank ulcer. | Ulcer, erosive gastritis, duodenitis, angina, ischaemic heart disease, biliary tract disease, oesoph. motility disorder, varices, stricture, surgery, H2 antagonists, lactose intolerance, anticipated poor compliance. |
| CARLSSON 1998 | 18-80 years; history of upper gastro-intestinal symptoms for ≥3months; episodes of upper GI symptoms occurring ≥2 days during the past 7 days. | Oesophageal ulcer or stricture, Barrett's oesophagus, peptic ulcer disease, history of oesophago-gastric surgery or GI bleed |
| GALMICHE 1997 | ≥18 years; heartburn as predominant symptom ≥3 months; Normal or erythematous oesophagus or non-circumferential erosive oesophagitis; heartburn on ≥2 days during last week of 2-week alginic acid run-in. | Erosive oesophagitis with significant erosions, ulcer, stricture, Barrett's oesophagus, complications of GORD, surgery, H2 antagonist or acid pump inhibitor use |
| HALLERBÄCK 1998 | ≥18 years; presenting with reflux like symptoms (at least 2 of following: retrosternal burning pain, epigastric pain, fluid acid regurgitation. | Oesoph. stricture, Barrett's oesophagus, ulcer, varices |
| HATLEBAKK 1999 | 18-80 years; heartburn as predominant symptom for ≥3 months and ≥3 days in past week during 14 days run-in. | Severe oesophagitis, Barrett's oesophagus, ulcer, gallstone, surgery, prokinetic or antisecretory drugs, alcohol misuse, concomitant disease, need for interpreter |
| RICHTER 2000 | ≥18 years; history of heartburn ≥12 months as predominant symptom of GERD and current episodes of moderate to severe heartburn on ≥4 of 7 days prior to endoscopy. | Conditions interfering with assessment of heartburn, ulcers, GI bleeding, Zollinger-Ellison syndrome, pancreatitis, malabsorption, IBD, severe illness, PPI, H2-antagonists, anticipated need for range of medications, including NSAIDs |
| VENABLES 1997 | ≥18 years; heartburn as predominant symptom for ≥3 months; patients who remained symptomatic after 2 weeks run-in. | History of ulcer, oesoph. stricture, GI surgery, severe illness, anticipated poor compliance |

In most studies, the duration of symptoms suffered by the majority of patients prior to recruitment was over one year, for four studies the duration is not clearly reported (Table 7.30).

**Table 7.30: Duration of symptoms at baseline (case study 4)**

| Study | Duration of symptoms at baseline |
|---|---|
| EGGLESTON 1996 | Unknown |
| BATE 1996 | Unknown |
| BATE 1997 | Mean circa 10 months |
| CARLSSON 1998 | Majority >1 year |
| GALMICHE 1997 | Majority >1 year |
| HALLERBÄCK 1998 | Majority >1 year |
| HATLEBAKK 1999 | Unknown, but inclusion criteria require minimum 3 months |
| RICHTER 2000 | Seemingly majority >1 year |
| VENABLES 1997 | Majority >1 year |

**Outcomes, analysis methods, and results**

Table 7.31 presents the outcome measures used by the studies included in case study 4. The rather crude outcomes used in the ESD were based on medication use alone; they did not directly match any of the outcome measures presented in RCTs. Also these were not homogeneous, as some measured freedom from (one or more) symptoms, and others the proportion of patients suffering a defined maximum of symptoms on study completion (e.g. <=1 day with mild symptoms within one week).

Given the brief abstract publication, Eggleston reported only percentages of outcome occurrences in each group of patients. Neither p-values or confidence intervals, nor statistical control of potential confounders were available from the publication.

For efficacy outcomes, the RCTs had on the whole used simple analysis methods and statistical testing for comparisons of outcomes between the different treatment and control groups. In addition, some had used longitudinal analysis methods for particular outcomes, or regression analysis methods to identify factors associated with particular outcomes.

Table 7.32 presents the results of outcome measures related to those used in the ESD by Eggleston *et al.* (1996). For a quantitative comparison, the odds ratio (OR) for Eggleston's results was assumed as 1.0, based on their statement: "in each group ~half of the pts received medication for 1mo" (data were requested from the authors but not made available). This was compared to a pooled OR of 4.94 (CI: 3.43-7.16). Figure 7.3 shows the results of cisapride vs. omeprazole in a L'Abbé plot.

**Table 7.31: Outcome measures used by studies in case study 4**

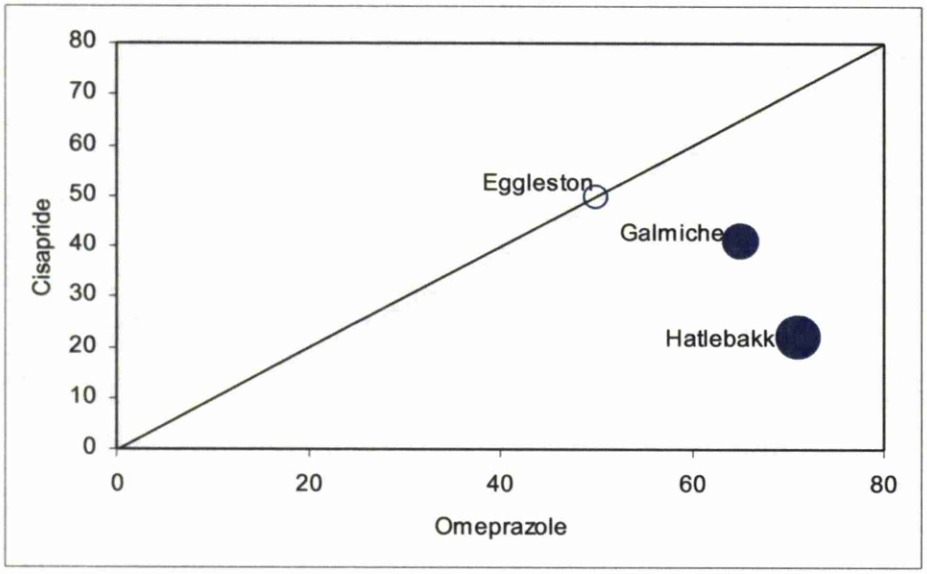
| Study | Outcome measures |
|---|---|
| EGGLESTON 1996 | *success rates* = reflux medication prescribed for <=3 months;<br>*failure rates* = switch of prescriptions in first 3 months or subsequent reflux prescriptions for >3 months;<br>*relapse rates* in 6 months in successfully treated patients |
| BATE 1996 | Proportion of patients without heartburn, regurgitation and other symptoms<br>Proportion of patients without symptoms<br>Odynophagia<br>Psychological well-being |
| BATE 1997 | Proportion of patients free of heartburn<br>Proportion of patients in remission of heartburn |
| CARLSSON 1998 | Complete relief of upper gastro-intestinal symptoms<br>Sufficient control of symptoms<br>6-month relapse rate<br>Quality of life |
| GALMICHE 1997 | Effective resolution (≤1 day with mild episodes of heartburn during last 7 days)<br>Complete absence of heartburn in past 7 days<br>Quality of life |
| HALLERBÄCK 1998 | Complete symptom relief (=all reflux symptoms ceased)<br>Improved (=fewer symptoms)<br>Responders (=satisfied with improvement) |
| HATLEBAKK 1999 | Adequate control of heartburn (≤1 day with no more than mild heartburn in 7 days before 4-week visit)<br>Antacid consumption<br>Severity and number of days with heartburn in 7 days before each visit<br>Severity of other (specified) symptoms<br>Adverse events |
| RICHTER 2000 | Complete resolution of heartburn<br>Proportion of patients with no heartburn on each day<br>No heartburn for 7 days prior to evaluation<br>Severity of other (specified) symptoms |
| VENABLES 1997 | Relief of heartburn after 4 weeks treatment (≤1 day of mild symptoms in last 7 days) |

**Table 7.32: Results of studies in case study 4**

| Study | Outcome measures | Results |
|---|---|---|
| EGGLESTON 1996 | "Success rates" (medication for ≤3 months) | OME: 75%; CIS: 70%; RAN: 79% |
| | "In each group, about half of the patients received medication for ≤1 month." | |
| BATE 1996 | Symptom-free after 4 weeks | OME20: 43%; Placebo: 14% |
| BATE 1997 | Patients in remission of heartburn at 4 weeks | OME20: 66%; CIM1600: 31% |
| | Patients reporting no heartburn in past 7 days after 4 weeks | OME20: 56%; CIM1600: 24% |
| CARLSSON 1998 | "complete relief" after 4 weeks | OME20: 41 (CI:34-48)%; OME10: 35 (CI: 29-42)%; Placebo: 19% (only endoscopy-negative patients randomised to placebo group!) |
| | "sufficient control of upper gastro-intestinal symptoms" after 4 weeks | OME20: 73 (CI: 67-79)%; OME10: 62 (CI: 55-68)%; Placebo: 35% (only endoscopy-negative patients randomised to placebo group!) |
| GALMICHE 1997 | Complete absence of heartburn in past 7 days at 4 weeks | OME20: 55%; OME10: 42%; CIS40: 29% |
| | ≤1day with mild heartburn in past 7 days | OME20: 65 (CI: 57-73)%; OME10: 56 (CI: 48-64)%; CIS40: 41 (CI: 32-49)% |
| HALLERBÄCK 1998 | Complete symptom relief after 2 weeks | RAN300: 41%; Placebo: 32% (no significant difference) |
| HATLEBAKK 1999 | Adequate control of heartburn at 4 weeks | OME20: 71 (CI: 64-78)%; CIS20: 22 (CI: 15-29)%; Placebo: 18 (CI: 12-24)% |
| RICHTER 2000 | Complete resolution of heartburn for a whole week at week 4 | OME20: 48%; OME10: 27%; placebo: 5% |
| VENABLES 1997 | ≤1 day of mild symptoms of heartburn in week 4 | OME20: 61%; OME10: 49%; RAN300: 40% |
| EGGLESTON 1996 | Relapse rates at 6 months | OME: 37%; CIS: 23%; RAN: 52%. |
| CARLSSON 1998 | 6-months relapse rate | 83% of 268 OME patients entering follow-up phase |

CIM=Cimetidine, CIS=Cisapride, OME=Omeprazole, RAN=Ranitidine

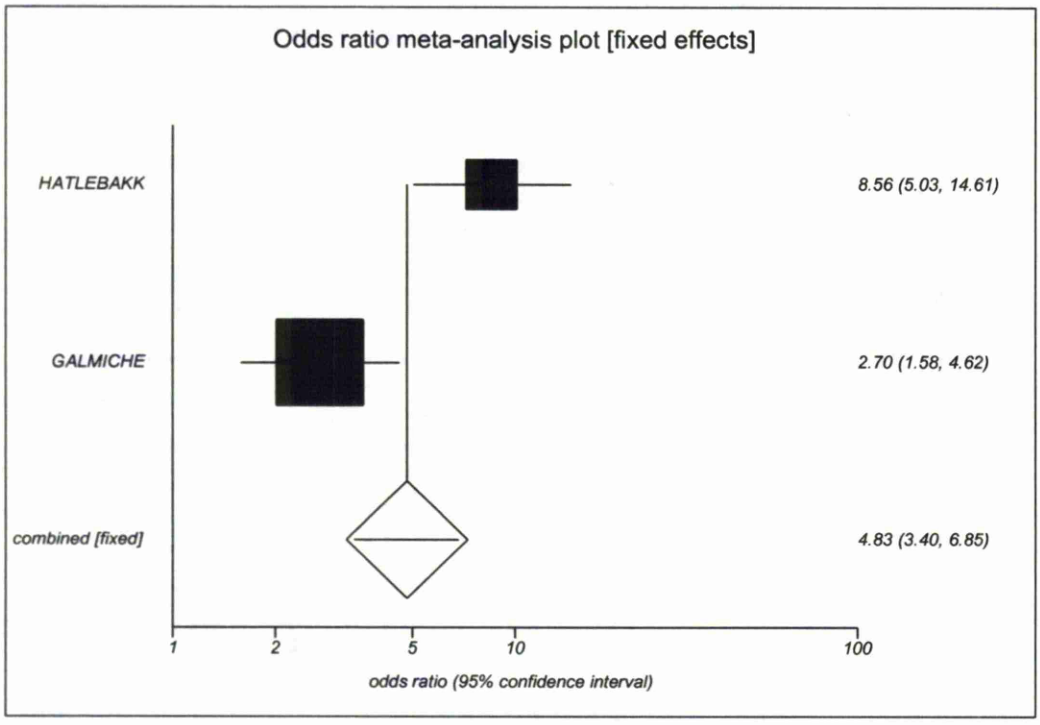
**Figure 7.3: Percent of patients with adequate control of heartburn after 4 weeks of omeprazole or cisapride**




| Outcome measures presented: | |
|---|---|
| Hatlebakk | <=1 day in past 7 with mild symptoms |
| Galmiche | No heart burn in last week |
| Eggleston | Medication for <=1months |

Notes: Numbers in Eggleston had to be estimated (see text).
Size of points is proportional to size of study population.

**Figure 7.4: Odds ratio meta-analysis plot of RCTs of 4-week outcome for omeprazole vs. cisapride, favouring omeprazole**




Note: ORs and CIs here are based on Mantel-Haenszel estimates.

**Discussion**

Comparability

The outcomes assessed in the ESD were not easily comparable with those from the available RCTs, mainly because Eggleston used a 3-month cut-off for defining treatment success, and moreover defined this by a lack of further treatment. The selection criteria of this ESD were also more restrictive than those used in trials, as only new cases of GORD were included. The samples could not be compared because baseline data were not available in Eggleston's abstract. Eggleston also reported no statistical adjustment for any possible baseline differences.

Comparison of effect estimates

Relative outcomes could only be compared between omeprazole and cisapride groups. From Figure 7.3 above it can already be gleaned that the effect size between those drugs was much larger (in favour of omeprazole) in the RCTs (Galmiche, Hatlebakk) than the ESD. As for absolute success rates, one might have assumed that Eggleston's outcome of "about half" of the patients in each group having received medication for ≤1 month means that ~50% of each group were completely free of symptoms after 1 month. If so, this compared reasonably well with the success rates achieved by omeprazole in the RCTs, particularly omeprazole 20mg. The dose of omeprazole used in the RCTs was either 10 or 20mg, however, a dose of 10mg is recommended only for ulcer relapse prevention (British National Formulary 2003), and it is therefore conceivable that patients included in Eggleston's study have used higher doses at least for their initial treatment. However, for cisapride and ranitidine the relevant RCTs quoted considerably lower success rates than 50% (Galmiche: CIS (40mg): 29%, Hallerbäck: RAN (300mg): 41% (after 2 weeks), Hatlebakk: CIS (20mg): 22%, Venables: RAN (300mg): 40%).

Given that the Mediplus database recorded prescriptions rather than actual medication use, it would have been reasonable to expect a greater rate of drop-out than might be observed in a closely monitored RCT. If so, the

outcome measure "discontinuation of prescriptions" would have over-estimated the drug effectiveness. It was interesting, however, that the absolute effectiveness estimates varied less between the study types than the relative success of omeprazole and cisapride. One possible explanation would be a differential dropout between the two drugs in the ESD (i.e. cisapride being more likely to be discontinued before symptoms are controlled).

The validity of the measurement of the intervention is also threatened by the practice of prescribing and using medications. It is possible that patients spread a month's medication supply over a longer period by using it less frequently than prescribed. The renewal of the prescription would not appear on the system for longer than expected. Thus when it does, it might mistakenly be considered as a new prescription and counted into the relapse rate.

Indeed the difference in relapse rates for omeprazole between Eggleston (37%) and Carlsson (83%) was remarkable. Both studies followed up successfully treated patients for 6 months. Obviously, Carlsson and colleagues followed patients actively by phoning them, and this could explain some if not most of the differential. A possible explanation is that Mediplus might have missed patients whose symptoms have reoccurred, e.g. if symptomatic patients did not seek further medication within 6 months. A further possibility would be that the patients were essentially different in both studies. Whereas Eggleston included only patients with a first diagnosis, most of Carlsson's patients had suffered symptoms for more than one year. Eggleston did not, however, report duration of symptoms.

Inclusion / exclusion criteria

All studies seemed to have sought to recruit patients with uncomplicated GORD. Eggleston and colleagues limited their sample to patients with a first diagnosis of GORD (it is unclear how far back the authors were able to verify the GORD-free status of included patients). Particularly with first

diagnoses, there is room for misdiagnosis, and affected patients may have stopped their GORD medication early and have been considered as successfully treated. It seems likely that the population sampled in RCTs was different, because it included patients with repeat-diagnosis of GORD. Thus the inclusion criteria of the ESD seemed more selective, which may seem surprising, but was an attempt to avoid the need for taking account of this particular baseline difference.

Eggleston used a limited set of exclusion criteria compared to RCTs. It is conceivable that some of the possible misdiagnosis in the ESD might have been avoided by excluding differential diagnoses. The exclusion criteria used in the RCTs varied inasmuch as some RCTs have excluded most conditions which could have caused symptoms similar to heartburn, whereas others have not explicitly done so.

Participation of patients and centres

Table 7.28 gives an indication of the number of centres involved in the various studies, which range from 19 to 106. Only Eggleston and Galmiche did not report the number of centres involved. Apart from Eggleston, three RCTs recruited patients in primary care (Carlsson, Hatlebakk, Venables), and Bate 1996 used both primary care and hospitals as recruitment source. Bate 1997 was the only study having recruited only from hospitals. The remaining three studies (Galmiche, Hallerbäck, and Richter) did not report the source of recruitment.

It is interesting to note that - probably due to the random sampling of eligible patients - the number of patients included in the ESD was amongst the smallest of all comparison studies; this may have been because the study was essentially undertaken as the basis for a cost analysis and therefore only a random sample of eligible patients was included. Together with most RCTs, Eggleston did not report the number of patients screened for inclusion.

Adjustment for baseline differences in ESD

The abstract by Eggleston did not report any adjustment for baseline differences, but a subsequent cost-effectiveness study by the same authors - which may have referred to this study - noted that the comparison groups were similar in terms of age, gender, and co-morbidity with other forms of gastrointestinal disease (Eggleston *et al.* 1998). It is of course not possible to say whether there were other confounders which related both to a physician's choice of first-line treatment as well as the outcome. Any perception of severity may have been a serious confounder, as omeprazole may have been chosen for more severe cases.

Follow-up

In this case study, the follow-up of patients was comparable between ESD and RCTs.

Contribution to effectiveness evidence

Eggleston and colleagues did not set a concrete study objective; rather they claimed to be undertaking an "hypothesis-generating exercise". They avoided any claim of providing either efficacy or effectiveness evidence, but concluded that "*there is little evidence to support a hypothesis of differences in response rates...*". Thus, the extraction of any effectiveness evidence from the study may have been unjustified. Interestingly, all comparable trials post-dated the ESD (previous trials mostly included different drugs or placebo control groups).

In 1998, a meta-analysis of the same three drugs investigated by Eggleston for GORD was published (Iskedjian & Einarson 1998); the study was funded by the Canadian branch of the same pharmaceutical company as Eggleston (which markets cisapride). The review concluded that omeprazole was the treatment of choice, but suggests that (an unusually high dose of) 80mg cisapride was most effective in mild cases. Four-week cure rates for cisapride were not analysed due to lack of raw data. The analysis has, however, neither assessed quality nor heterogeneity. A meta-analysis of

treatments in more severe GORD concluded that proton pump inhibitors were more effective than H2-receptor antagonists and cisapride in terms of their overall healing proportion: 83.6% [CI: 79.2-88.1], 51.9% [CI: 46.9 - 56.9] and 37.9% [no CI reported] respectively (Chiba *et al.* 1997). However, comparisons were made between rather than within studies. The later Cochrane review (van Pinxteren *et al.* 2001) concluded that "*when patients are selected primarily based on symptoms (i.e. heartburn meeting certain criteria) and diagnostic probability of GORD is high, proton pump inhibitors are superior to both H2-receptor antagonists and prokinetics in achieving heartburn remission*". PPIs are now firmly established as first-line treatment for GORD (National Institute for Clinical Excellence 2004). Thus, the hypothesis of equal response rates raised in the Eggleston study was not confirmed subsequently.

Eggleston and colleagues stressed the fact that their data were primary care based, rather than having assessed treatments in specialist centres which is usual for clinical trials. Indeed, the three RCTs undertaken in primary care were published after Eggleston's study (Carlsson, Hatlebakk, Venables). However, GORD severity may be more likely to account for any difference in findings, than treatment setting. Galmiche's RCT suggested that the relative efficacy of omeprazole vs. cisapride was higher in the presence of oesophagitis.

Outcomes assessed by GORD trials vary, and the Cochrane review authors have criticised the lack of comparable and meaningful dichotomous outcomes, such as cure rates (van Pinxteren *et al.* 2001). This means that several studies have been excluded from both the Cochrane review as well as this case study on the basis of their outcome measures. Arguably, cure rates are meaningful, since if no total symptom relief is achieved, treatment switches or further investigations are initiated in practice. This is highly relevant in terms of determining the overall costs (to patients and the health service) of different treatment options.

Eggleston's cost analysis of the three comparison drugs (Eggleston *et al.* 1998) argued that on the basis of the 6-month cost being cheapest for cisapride, this should be the first-line treatment of GORD. However, this cost study did not mention patient outcomes and assumes equivalent effectiveness of the three drugs, based on the very ESD, despite having claimed there that this was merely a hypotheses-generating exercise. Referring to the ESD, the background section of the cost study concluded by saying *"given such therapeutic equivalence it is appropriate for this analysis to concentrate on identifying the strategy that achieves this common outcome with the lowest consumption of scarce health care resources."* (Eggleston *et al.* 1998 p.14) However, the cost study and its conclusions (suggesting a step up approach of treatment) were in contradiction with several other publications and have been criticised by an evidence-based healthcare review (Bandolier 2002).

**Concluding comments**

In summary, the database study by Eggleston provided probably more confusion than enlightenment to the body of effectiveness evidence of GORD treatments. Whereas it claimed to be hypothesis generating, the authors have based a cost analysis in its findings. In the absence of any consideration of possible confounding by indication or selection bias, presenting a primary care-based assessment (as opposed to specialist care) seems a comparatively minor contribution. A full manuscript has never been published.

## 7.6 Discussion

In the four case studies, it was attempted to match ESDs with existing RCTs previously included in reviews by the Cochrane Collaboration. The comparison of results between ESDs and RCTs did not show any systematic difference in terms of effect estimation. This is perhaps unsurprising, given the findings of recent systematic reviews of randomised vs. non-randomised study designs (Britton *et al.* 1998; MacLehose *et al.* 2000). There was no indication that ESDs over-estimated the effectiveness of an investigational drug, relative to the RCTs' conclusions.

One may assume that ESDs outweigh some of their potential methodological shortcomings with a larger sample size and longer follow-up. In the four case studies, most ESDs were considerably larger than comparable RCTs. Particularly those ESDs based on the NRMI-2 and GPRD belong to the largest identified (Tiefenbrunn, van Staa). By contrast, Eggleston's was a very small ESD, particularly given that the source database (Mediplus UK) would have permitted a much larger study.

Case study 1 illustrated the diverse nature of ESDs: whereas van Staa included nearly 16,000 patients from the GPRD, Sebaldt reported on a 3-year follow-up of a mere 61 patients in a single centre - still more patients than were included in comparable RCTs. Nevertheless, these two examples demonstrate the potential problems with these two extremes very well. On the one hand, a cottage industry of single-centre databases with little outside involvement or quality control and potentially poor broader representativeness probably adds little to a body of effectiveness evidence. On the other hand, a large non-disease specific database such as the GPRD or Mediplus present inevitable limitations in data quality, definition and completeness.

The definition of outcome variables is critical to an ESD. This, however, is inevitably very challenging for effectiveness studies of databases which were essentially designed to identify adverse drug reactions (such as the GPRD). However, this is not to say that subject-specific databases are necessarily any better equipped with good quality outcome variables (as the ERCF case has shown).

In terms of follow-up, both ESDs in case study 1 offered follow-up periods of up to around 3 years, which exceeded those of comparison RCTs (max. 2 years). Case studies 4 and 2 involved suitably short follow-up periods in all included studies. However, in case study 2, the ESD had to restrict its sample selection based on the availability of a suitable follow-up period for each patient. Patients transferred within 48 hours of having received the intervention had to be excluded to avoid missing outcomes data. The authors were, however, able to offer a comparison between patients who were transferred early and those included in the study. The ESD in case study 3 differed from included RCTs not only in terms of a longer follow-up of the ESD, but also inasmuch as it observed intermittent use of the interventional drug over 10 years, rather than continuous use as was assessed in most RCTs. The only RCT of comparable duration to the ESD was rather different in the nature of its comparison treatments. Thus, ESDs did not offer outstandingly longer follow-up than comparison RCTs. However, it may be argued that with ongoing data collection in longitudinal databases, the available follow-up periods will increase in future.

Another point on which ESDs are often defended is their likely inclusion of a broad range of patients not normally represented in RCTs. The ESDs included in the four case studies offered little direct discussion of this point. Eggleston included so few patients of the large Mediplus UK database that it is questionable that their sample was representative of primary care practice in the UK. Similarly, McDougall was based on data from only one tertiary centre, which may not have been comparable to other centres or primary care. By contrast, van Staa's study based on the

GPRD included all primary care patients prescribed the investigational drug of several dozen practices across the UK, which is arguably one of the most convincing sources of data from everyday clinical practice.

The database used by Tiefenbrunn also included data from some 14% of US hospitals, but despite efforts to represent normal clinical care (e.g. by excluding patients enrolled in the large GUSTO trial) the database was somewhat biased towards hospitals with coronary interventional facilities (Rogers *et al.* 1994). This probably mattered less for a study testing an interventional cardiological therapy, as non-interventional centres would automatically be excluded. The generalisability to a naturalistic setting is primarily determined by the source database, and as more and better quality routinely-collected data become available, for example through the UK NHS' Connecting for Health information technology programme, ESDs could become more able to reflect treatment use and outcomes in everyday practice. Having to fall back on prescription use as an outcome measure will hopefully become a thing of the past.

The representativeness of the sample compared to a desired target population can be strengthened by broader inclusion criteria and assessed through reported baseline characteristics. Here one would expect ESDs to include more severely ill, maybe multi-morbid, older patients than RCTs. Indeed van Staa (case study 1) and Tiefenbrunn (case study 2) included more women than comparable RCTs, van Staa included older patients, and Tiefenbrunn more patients with previous AMI. McDougall (case study 3) included patients with far longer disease history (pre-study entry) than did comparison RCTs.

This chapter also explored issues regarding the potential contribution of ESPDs to a body of effectiveness evidence. Two questions may be asked: (1) is it possible to control alternative explanations for the findings of an ESPD (in particular bias and confounding), and (2) can ESDs provide evidence for causality?

The case studies identified typical sources of observation bias in ESDs, in particular detection bias, and attrition bias (caused by differential dropout rates between study groups). In addition, longitudinal databases may be subject to bias through the use of historic controls, cohort, age, and survivor effects. Detection bias is minimised in investigations of unanticipated or unambiguous outcomes (such as mortality). This is not always the case in ESDs. Attrition bias is also not always possible to account for in a retrospective database. These biases potentially lead to over-estimation of treatment effects, whereas non-differential under-reporting may lead to under-estimation.

There are several statistical methods available for adjusting for confounding, and together with sensitivity analyses, these may reassure readers that sufficient control has been applied and results are robust, particularly if valid and reliable measures of all known confounders were available.[5] However, adjustment is no guarantee of full control of all possible confounders. Thus, some threats to validity are impossible to control, even with perfect data quality.

Whereas the question of whether *sufficient* control through adjustment is possible is beyond the scope of this thesis, the case studies support suggestions from the existing literature that congruence with RCT results is frequently the case. However, comparisons were complicated by differences in outcomes, selection criteria, and follow-up.

On the question of positive evidence for a causal relationship measured on the criteria of chronological relationship between cause and effect, the strength of a relationship, a dose-response relationship, consistency, and specificity (Elwood 1998), ESDs offer considerable potential, at least in

[5] To the average reader, the quality of the adjustment may be secondary to it having taken place, as the sophistication of adjustment methods has surpassed the level of his/her understanding. A publication in a reputable journal may be seen as sufficient justification to take the results seriously.

theory. The chronological relationship between drug intervention and outcomes is usually easily established, particularly where ESDs are also in a position to determine previous exposure to the investigational drug. If under-reporting (resulting in under-estimates of the true effect) is indeed the main validity problem, a highly significant effect can provide strong evidence. Comprehensive databases should be able to assess possible dose-response relationships as patients are receiving the investigational drug over varying periods and in varying doses. Finally, size of and variability in the sample may provide for the assessment of consistency and specificity of results. Thus, ESDs should aim to provide evidence for causality in practice by demonstrating how they meet these criteria.

### 7.6.1 Methods and limitations

The incompatibility of outcome measures limited the comparability in some case studies. For example, of 17 trials included in the Cochrane review by van Pinxteren *et al.* (2001), nine could not be included in case study 4, because they used outcome measures not comparable to those used in the ESD. Outcome measures used in trials were clinical parameters and measurements which the prescription database used by the ESD (Mediplus UK) would not have been able to collect.

Strictly speaking, no case study fulfilled all the criteria for comparability (outcome measures, population, intervention, selection criteria, adjustment for confounding in ESD), and comparability is at times difficult to judge due to limited data presentation.

The comparison RCTs were exclusively drawn from existing Cochrane reviews. Whereas this meant a timely close proximity with the ESDs, newer trials might have been missed. Similarly, trials excluded by the Cochrane reviewers could have been missed. This seems possible in theory in case study 3 (amongst the studies excluded by the Cochrane review are four further potential candidate RCTs [excluding continuations of earlier trials],

but they are early trials of analgesic treatments and unlikely to include comparable outcomes).

There are limitations to the usefulness of such comparisons, since they test the validity of an ESD against the standard of an RCT. It might be argued that the validity (external, but also internal) of an ESD cannot be measured satisfactorily against an RCT, because the nature of the hypothesis *should* differ. Thus, one is looking for an indication that an ESD assesses a distinctly different hypothesis, e.g. one aimed at effectiveness rather than efficacy. A meaningful difference between both study designs (in terms of generalisability to a naturalistic situation and defined target population, and follow-up) thus should be expected.

For the assessment of the quality of ESDs, a checklist was used, which was based on instruments designed for the quality assessment of non-randomised studies as well as suggestions from the literature. The checklist focused on internal and external validity, data processing, analysis, and reporting. This currently rather lengthy list offers a basis for the development of an assessment instrument suitable for ESDs. In both quality assessment and data extraction, it would have improved the validity of this review, if a second researcher had been available to double-check extracted data and rate the quality of the studies.

Lastly, only very limited numbers of outcomes were compared for a small number of case studies, and therefore the findings cannot be considered conclusive.

## 7.7 Concluding remarks

The four comparison case studies offer no support to the claim that observational studies tend to over-estimate effectiveness. Their actual comparability was limited due to differences in outcome measures, or patient selection criteria, which may explain some of the differences in results. The included ESDs were not always clear about their hypothesis and thus their proper contribution to effectiveness evidence. In retrospect, none of the ESDs seem to have been seminal studies in building the relevant evidence base.

ESDs presented considerable scope for biases and may have been unable to control fully for confounding by indication. They were essentially plagued by the pre-existing nature of the databases used. This meant that they had no control over data collection or data quality, except through adapting their own study design to accommodate known limitations. However, complete adjustment for confounders may be an illusion, and as a minimum, the extent of uncontrolled confounding should be estimated.

There is an urgent need for stricter quality controls on the conduct and reporting of ESDs from all sources, particularly industry, as the potential for misuse and misinterpretation of data and analyses is currently far greater than necessary, given even the existing methodological knowledge. More methodological research is needed to improve the validity of ESDs, bearing in mind that RCTs may not always be the best test of validity. The goal has to be to improve confidence in ESD results particularly where no RCTs are possible.

# 8 Discussion

This chapter draws out the methodological findings from the Dornase Alfa Case Study and reviews presented in previous chapters, and highlights how they have contributed to the existing body of knowledge; the chapter falls into the following sections:

8.1 Databases for ESDs
8.2 Design and analysis issues for ESDs
8.3 Generalisability
8.4 Data protection
8.5 Next steps for ESDs

Much previous work has focussed on databases, rather than effectiveness studies based on the analysis of their data. This thesis clearly presents a further step in as much as it explored the detailed methodological issues incurred in the use of databases for effectiveness evaluations of drug treatments.

## 8.1 Databases for ESDs

The Review I have undertaken suggests a recent rise in actual publications of ESDs in high-quality peer-reviewed journals (those included in the Abridged Index Medicus). The first identified study was published in 1989, and nine of the included 42 studies were published in 2001 alone (the last complete year included in the search). This underlines the importance of methodological efforts to improve the quality of such studies. In this section, I focus on the role of databases in this endeavour. The main attention has to be on improving data quality and completeness, reducing opportunity for bias, and improving the control of confounding by indication.

The vast majority of studies identified in the Review were retrospective and used pre-existing databases. Some of these were disease-specific and may or may not have been collated with particular research hypotheses in mind. Other databases used were generic large data sources collated for different purposes such as claims management or primarily for the identification of adverse drug events (such as GPRD). Such generic data sources inevitably lack variables needed for a comprehensive analysis of a particular hypothesis; authors have at times resorted to the use of approximate measures (such as occurrences of prescriptions). This invariably results in potential biases. However, even disease-specific databases do not guarantee the availability of well-defined and high quality variables, as the Dornase Alfa Case Study demonstrated. This was despite the fact that this was essentially a post-marketing database for dornase alfa, and relevant specific research hypotheses would have been anticipated.

Therefore, the quality control of key variables is absolutely vital in databases. The ERCF collected a vast number of variables; only some were

subject to routine quality control processes. For a small number of key variables, we were able to undertake a data validation against existing clinic information. Enthusiastic database initiators are understandably keen to include a large number of variables to test a range of foreseeable (or un- foreseeable) hypotheses later. However, for effectiveness hypotheses, and particularly where only a moderate effect size has to be expected, attention to data quality must be given priority over including a large number of variables.

The availability of variables in the source database is a key determinant of the success of an ESD. Where studies are forced to use proxy measures for exposure, such as prescription records, the exposure measurement loses precision. In some reviewed studies, even outcomes were assessed through prescriptions rather than clinical parameters, for example in Eggleston's study of GORD, and several others assessing antibiotics in acute infections (Bowman, Huang, Lawrenson). The validity of such outcome measures depends on how closely a new or alternative prescription represents "treatment failure" (and no new prescription represents "cure"). This would depend on the inevitability of treatment (e.g. non-drug alternatives may be used but not recorded), the likelihood of patients to seek treatment, and to seek it at the same source, and the required follow-up period (and danger of differential follow-up, e.g. due to side effects). Non-disease specific databases are unlikely to have recorded the indication and rigorous outcome variables for many treatments they may record. This severely limits their potential use in ESDs. Any databases should ensure that the link between diagnosis / indication and treatment can be clearly established for subsequent analysis.

This implies that we should be concerned about the use of non-treatment specific databases in the evaluation of drug effectiveness, since their scope for sufficient attention to data quality will be severely limited - notwithstanding their possible lack of appropriate variables in the first place! This is, however, what is proposed by some in the HTA community

(Rawlins 2008). This does not negate the value of such data sources for HTA, but their potential to generate effectiveness evidence for drug treatments may still be limited. Similarly, the NHS "Connecting for Health" IT programme may be raising hopes one day to be able to use vast amounts of routinely collected data for outcomes research on a population basis. In the case of evaluating most drug treatments, particularly on what are often moderate or long-term effects, this hope may be considerably misplaced.

Having just emphasised the importance of data quality, it was both surprising and disappointing that many reviewed reports did not describe data quality in any meaningful detail. Given the multitude of data quality problems identified in the ERCF (despite the detailed data quality assurance processes in place there), this is one of the most serious shortcomings of existing ESDs identified by the Review. Since this work was undertaken, more guidance on the conduct of databases has become available, notably through work by the Agency for Healthcare Research and Quality (Gliklich & Dreyer 2007). However, guidance on reporting ESDs already existed and was clearly not heeded. Therefore, more effort needs to be invested in the short-term into ensuring that databases - and particularly those subjected to ESDs - conform to existing guidance, and can objectively demonstrate compliance. The DoCDat directory in the UK may present a resource for supporting such assessments. It currently assesses the quality of databases registered on DoCDat, but this assessment relies on interviews and does not seem to be verified.

The Review noted that some databases have already been used for several publications of drug ESDs, sometimes by the same or a similar group of authors. The setting up of an ESD involves much work in terms of negotiating access, validating and processing data, and exploring possibilities for analysis. Therefore, it seems likely that once these preparatory stages have been passed through successfully, several studies are run on the same database by the same group of researchers. It remains

unclear whether many databases are never used for effectiveness studies, or to what extent they may have been subject to failed attempts of undertaking such studies. None of the studies identified used any of the UK databases registered by DoCDat. Raftery *et al.* (2005) also identified a minority of studies having used UK databases for effectiveness research, usually evaluating non-drug treatments.

This raises the issue of access to databases for research / HTA. As the Dornase Alfa Case Study showed, there are many players who have interests to protect and who may influence the decision of whether to undertake an ESD or give permission for such a study: pharmaceutical companies, the organisation funding the database (which may be a pharmaceutical company), staff responsible for the database, and the clinics providing data which may give insights into variations in patient care. In the case of the ERCF, clinicians and researchers responsible for the database had a strong interest in protecting its credibility. This was compounded by the fact that further funding for the database was being sought on the basis of arguing for its use as a research tool. Many ESDs were published by researchers involved with the database and/or regularly involved in analysing it, suggesting that the use of databases may remain limited to a core group of individuals. Apart from wishing unfavourable outcomes not to become identifiable, clinicians may prefer data not to be made available to outside research teams, as they may wish to conduct research themselves. A current example is the bariatric surgery registry - a collaboration between professional organisations and a commercial company - which is unwilling to open the database to NHS researchers.

Whereas this did not apply in the ERCF, it is possible that patients may also prefer certain hypotheses not to be addressed, as they may risk losing a favoured treatment option. This may be particularly relevant when cost-effectiveness evaluations are proposed, as in the case of lysomal storage disorders, where both patients and clinicians are currently actively undermining such a study. Database funders and operators should clarify a

priori the access conditions and rights for outside researchers. This process needs to take account of consent from contributing clinicians and patients if necessary (see later section on data protection).

It was not possible in the Review to assess the magnitude of any publication bias, albeit that such a bias is highly likely. A majority of studies presented significant findings, and whereas overall not many studies were directly funded by the pharmaceutical companies marketing the investigational drug, no study showing negative results for an investigational drug had received direct funding from the relevant company. Of course, the issue of publication bias is more complicated in the case of ESD: there may be a bias towards publishing significant results, which could be significant in either direction. However, if the results do *not* concur with those of randomised trials, these studies are often criticised for confounding by indication, and this may present a barrier to publication. In the UK, the DoCDat directory of databases reports publications from registered databases but is not complete and does not record non-published analyses. If it did, it could theoretically assess publication bias at least of UK-based ESDs (if any were indeed undertaken on UK databases!).

The Dornase Alfa Case Study identified some very specific lessons for routine data audits, data cleaning and data verification processes. Comparisons with original records, such as in the DQR, are important in estimating the extent of possible errors. Databases should therefore plan such verification processes from the outset and should ensure that they are undertaken periodically. Similarly, routine data checks can be pre-planned and supported with appropriate variables for cross-referencing. Data audits need to include routine comparisons between centres to identify and subsequently assess any systematic differences between data from different centres.

However, researchers also need to understand the entire process of data collection before they can make assumptions about the appropriateness of

the data for any research hypothesis. Moreover, particularly in multi-centre databases or for less clearly defined variables, data collection processes may be subject to assumptions and interpretations by people from a range of different perspectives (clinicians, database managers, data entry clerks, and not least researchers). Examples of resulting differences are the choice of lung function results reported by different centres in the Dornase Alfa Case Study, as well as the rather loose interpretation of what constituted "exacerbations". Database operators should therefore give careful attention to standardisation of data definitions and assess their reliability prior to implementation. Also, the Dornase Alfa Case Study identified that the motivation of contributing centres to contributing high quality data could have been improved. Incentives could be made dependent on data quality, particularly where contributors receive payments.

## 8.2 Design and analysis issues for ESDs

The Dornase Alfa Case Study and Reviews have identified a number of problems in the conduct and analysis of ESDs which have implications for their validity. The main challenge for ESDs is to avoid a range of possible biases (information bias, selection bias) and confounding; hence careful study design and analytical approaches are required to identify potential bias and control it. This section brings together the discussion of specific problems identified as part of the work on this thesis.

The primary criticism of ESDs concerns their susceptibility to bias due to a non-randomised treatment allocation. In many of the reviewed ESDs, as well as the Dornase Alfa Case Study, this led to a range of baseline differences between comparison groups (except of course for those variables used for matching by some studies). Most studies have used analysis methods aimed at adjusting for such differences as well as other known confounders. However, many reports were not detailed enough to permit any assessment of the adjustment process; the reporting of both adjusted and unadjusted results was rare, and variables used for adjustment were not always listed.

Only one study identified in the Review used the propensity score method, which is now becoming more widely used for such analyses. A propensity score is a conditional probability that a subject will be "treated" based on an *observed* group of covariates (Rosenbaum & Rubin 1983; Newgard *et al.* 2004). However, a recent review suggests that propensity score methods so far have not shown themselves to yield substantially different estimates compared to conventional multivariate analysis methods (Stürmer *et al.* 2006). There are other methods available (Berger 2008) and the assessment of their respective strengths will require further work. In any

case, researchers need to not only report but also assess the extent of any adjustment and residual confounding.

In many reviewed studies, as well as the Dornase Alfa Case Study, very little was known about the rationale for the initial treatment allocation. In the Dornase Alfa Case Study, clinicians held that - besides licensing indications - financial considerations determined how many patients could be initiated on dornase alfa. Eventually, patients at most centres were commenced on treatment following a successful trial period; non-responders did not continue treatment beyond this period. Thus, the treatment choice could have been influenced by a variety of factors unknown to the analyst and/or been difficult to measure and thus control. Only one ESD explicitly reported having adjusted for variables which the clinician on the team routinely used for his treatment decisions (Choi). Not being able to control for such factors poses a serious risk of selection bias, which may be unmanageable in retrospective analyses. Database designers therefore need to endeavour to record such variables.

The measurement and recording of variables in a database are likely to be less rigorous than under trial conditions. Some reviewed studies have reported regular calibration of their measurement instruments, but this may not be the norm, even less so across several centres. However, a beneficial feature of an ESD may be that blinding to the research hypothesis may be less necessary, where outcome assessors may not be aware of future effectiveness evaluations using their data. Nevertheless, it is likely - particularly in single-centre clinical databases - that clinicians have in mind particular research hypotheses which they want to address in the analysis of their database. Thus, observation bias is still a real possibility. In one study (McDougall), one clinician had scored the global assessment of all patients. He died before the study was conducted, but it is possible that he foresaw such use of his data. How likely such bias is will not only depend on the preference of the assessor, but also on the degree of subjectivity of the measurement process. In McDougall's case, several

outcome measures chosen were subjective scoring instruments, for which no validity or reliability data were reported. Database designers and researchers therefore ought to consider testing the reliability of instruments in a given data collection context.

The definition of variables poses a variety of potential threats to validity and precision. Firstly, an exposure status to the treatment of interest needs to be assigned to each individual. The Dornase Alfa Case Study highlighted several possible pitfalls in this regard. Some centres did not record treatment trial periods on dornase alfa, which means that some patients had received the treatment at a point unknown to the analyst and would have been categorised as non-users. This indicates a lack of consistency in variable definition and data collection. Similarly, the categorisation into dornase alfa use groups is open to criticism. Short-term treatment and interruptions of treatment could have been handled in a number of ways. For example, an intention-to-treat analysis could have compared all non-users (recognising that there was room for misclassification of those also) with patients having ever tried or used dornase alfa. However, the distinction into continuous and intermittent use offered the theoretical possibility of detecting a dose-response relationship as well as generating a group better comparable to continuously treated trial patients. The lack of reliable data on treatment compliance observed in the Dornase Alfa Case Study (and indeed in reviewed ESDs) needs to be considered in this context also. Taken together, these findings illustrate the considerable complexity in defining the exposure variable alone. However, it can be argued that a smaller number of well-defined variables collected in a database is more useful than a large number of poorly defined and controlled variables.

The Dornase Alfa Case Study undertook very detailed data cleaning and verification processes. It is hard to over-estimate the importance of this work, and analysts need to allow sufficient time for a detailed assessment of the quality of variables they wish to use in the analysis. Where under-

reporting seems likely (particularly of outcome variables or drug exposure variables), such as in the case of deaths and exacerbations in the Dornase Alfa Case Study, analysts at least need to assure themselves that the level of under-reporting is the same across all comparison groups.

Data may be missing for a variety of reasons, which represents another source of bias. The Dornase Alfa Case Study identified that different centres saw their patients at different intervals. Hence, mildly affected patients may not attend frequently enough to generate a minimum number of database reports necessary for analysis (this may have been the case in Johnson's study). Similarly, patients not continuously enrolled with a claims database cannot easily be included in studies (the CF patients for whom reporting continuity was lost on transfer between centres may have been excluded from certain analyses, if this had not been identified). This alone can give rise to selection bias, as transferred patients or less stable populations may systematically differ from others.

Misclassification is of course also possible for outcome variables. Outcome variables, however, pose additional threats to validity, if assessors are not blinded. ESDs do not comment on blinding, but in several reviewed studies, one may suspect that clinicians contributing to a database would have been aware of the later researched hypothesis. In addition, a doctor's perception of a treatment may still distort his or her assessment of patients. Similarly, there may be detection bias at play, where doctors are aware of possible consequences of treatment or more aware of complications in patients under treatment and thus more likely to detect them there (e.g. fractures in patients treated with etidronate as in case study 1).

A degree of misclassification of intervention or outcomes is to be expected in databases, not least due to missing claims or under-reporting. Misclassification of outcomes seems most likely where their measurements are subjective, data linkage has to be relied on, or where assessors may be

aware of the study hypothesis. Misclassification of patients to intervention or control groups seems most likely where claims are missed, or where medication might be sourced outside the observed system. Both would underestimate the effectiveness of the treatment. Given the many uncertainties in terms of variable definition and measurement, it is surprising that only few reviewed studies reported having undertaken any form of sensitivity analysis.

A possible source of error or even bias may stem from not being able to secure baseline measurements, e.g. where patients are recruited onto a database *after* having first tried or commenced the interventional drug. Even where sampling starts at the licensing date of the drug, patients may have participated in pre-marketing trials, particularly if they are suffering from a rare disease treated in specialist centres (such as CF). Several reviewed studies aimed to exclude patients who were not clearly free of the use of the drug for a significant time before the beginning of the observational period. Ideally, a database would determine previous use of drugs of interest. This method has not been reported in any studies, but would be helpful for excluding patients who started the drug early during their enrolment on a database.

Reviewed studies did not describe their choice of variables in detail. In some cases, it might be suspected that several outcome variables might have been available, but there was no rationale presented for choosing those to be analysed and ultimately included in the published report. This could hide serious biases, and it is difficult to see how such bias could be detected, unless authors clearly state all outcome variables available and the basis for selecting any for inclusion in the analysis. In the case of dornase alfa, the treatment outcome in terms of number of exacerbations may be of at least as much interest as lung function changes. Johnson et al. (1999) have not reported this outcome at all, although it would have been measured. However, if the ESCF is similar to the ERCF in its multiple

recording of exacerbation events, then it is easy to see why Johnson and colleagues may have had to ignore that outcome.

The duration of available follow-up may make a database attractive for effectiveness research. The Dornase Alfa Case Study has highlighted a number of further issues relating to longitudinal analyses of chronic disease populations: the known difficulty in isolating the effect of aging on lung function decline from an effect of treatment, comparing patients who have had different experiences of treatment by virtue of their age cohort, as well as coping with differential survival of severely and less severely affected patients. As a database ages, representativeness and/or continuity of follow-up may be affected. Similar problems might have affected the longest ESD in the four comparison case studies (McDougall). Thus, the main advantages of ESDs cannot be taken for granted, but necessitate careful assessment, study design and analysis so as not to turn into major pitfalls.

The databases used in the identified studies are to a large extent longitudinal data sources, some of which are continuing to collect data and will thus provide increasingly long periods of follow-up. A long follow-up is sometimes seen as an advantage of ESDs. Nevertheless, only some seven studies (16%) included a follow-up of more than three years on average. In many cases a hypothesis does not require particularly long follow-up, e.g. in the evaluation of acute treatment or some preventive interventions, such as vaccinations.

Twenty-two of the reviewed studies addressed interventions for chronic diseases and conditions, including HIV/AIDS. It is these for which the possibility of a longer follow-up period is of interest. However, the Dornase Alfa Case Study highlighted the limitations of longitudinal analyses due to cohort and survival effects. Similarly, excessive duration of follow-up may introduce selection bias threatening external and internal validity. Small databases, however long their follow-up, may not support the

complex design and analysis techniques required for control of confounding in effectiveness evaluations. There is thus currently limited experience with significantly longer follow-up analyses than would be realistic for RCTs.

Rawlins (2008) recently argued for historical controlled trials (essentially cohort studies which compare groups treated - or not treated - at different time periods), as well as prospective data collection on particular disease cohorts with a view to subjecting data to such study designs later on. He suggests that key Bradford-Hill criteria for causal association could be used to judge the suitability of such study designs (Hill 1965). It can indeed easily be argued that ESDs have considerable potential to meet these criteria and hence theoretically are able to assess treatment effectiveness in the absence of randomisation. However, this does not address the considerable data quality issues identified in the Dornase Alfa Case Study, as well as the potential cohort and period effects affecting effectiveness analyses in chronic disease patient populations (such as CF).

## 8.3 Generalisability

Generalisability is an often-quoted advantage of observational studies over RCTs and therefore warrants closer scrutiny here. The essence of generalisability is *abstraction* from the specific situation of the investigated sample to a general situation (Rothman & Greenland 1998). Generalisability of an individual study has to be judged based on an understanding of the sampling process, as well as the research hypothesis in question (Hennekens & Buring 1987).

How much different types of hypotheses (of efficacy or effectiveness) can be at odds with each other is illustrated by this extract from an epidemiological text on intervention studies (Hennekens & Buring 1987, p. 204):

> *"Consequently, any procedure that maximises compliance, thus increasing the chances of obtaining a valid result, will positively affect the ability to generalise that finding to other populations."*

It is implied here that applying stringent controls on compliance can increase the internal validity of an intervention study. This is true for studies assessing efficacy, not, however, for effectiveness. Effectiveness relates to the outcomes observed in populations (or subgroups) - usually outside trial conditions. Here, (non-)compliance is not a factor to be influenced and controlled, but rather one to be observed and taken into account. Thus arguably the (internal) validity of a study assessing effectiveness - and thus also its generalisability - will *increase* if factors such as compliance are left un-influenced (but not un-measured) by the study design.

Frequently, hypotheses are investigated on a defined sample of subjects taken from a target or reference population. The generalisability of the study of an efficacy-related hypothesis will benefit if the sample can be clearly distinguished as having the condition of interest (and preferably no other), and if the exposure to the intervention is rigorously controlled (i.e. patients are either receiving the drug or not). If so, it should be possible to infer from the results, whether the intervention has the potential to cause a certain outcome in patients with particular characteristics. In investigating an effectiveness-related hypothesis, the statistical representativeness of the study sample in relation to the target population might be argued to be more important. Here researchers aim for generalisability of their results to a typical patient population in real life (which probably exhibits a mixture of compliance patterns, co-morbidities etc.).

Very few reviewed studies discussed generalisability. The characteristics of excluded patients were hardly ever reported, and in many cases no indication of the representativeness of the source database or the sample in terms of any target population was given. The sample selection from the source database was usually achieved by applying a set of inclusion and exclusion criteria. An examination of these, together with the documentation of baseline characteristics, should permit limited judgements on the likely representativeness of the sample (GAO 1992; Rittenhouse & O'Brien 1996).

However, exclusion criteria used by the ESDs centred on issues of data validity, i.e. excluding possibilities for misclassification or erroneous judgments on key variables. This included typical criteria such as previous treatment with substances similar to the study drug, but also administrative criteria such as excluding patients with missing data on drug dose, or other missing data (albeit that few studies report such exclusions). Thus patients lost to follow-up or with incomplete or insufficient follow-up (e.g. due to leaving the database) were excluded a priori. If available

duration of follow-up is used as an exclusion criterion, the resulting sample could easily vary systematically from patients with shorter follow-up. Similarly, some studies excluded transferred patients (e.g. Tiefenbrunn *et al.* 1998; Peterson *et al.* 1999; Krumholz *et al. 1998, 2001*), and may thus have jeopardised generalisability. The main point here is that incomplete reporting in all or any subgroups can seriously jeopardise the potential for an ESD to be representative and benefit from a database's wide recruitment base (even if that itself is representative).

Whereas baseline differences between comparison groups were readily presented, authors rarely reported characteristics of excluded patients, such as in the study by Tiefenbrunn *et al.* (1998). However, this would have enabled at least some assessment of the representativeness of the sample. It is encouraging that disease severity or co-morbidities rarely featured as exclusion criteria. This may support the argument that ESDs have comparatively high external validity (GAO 1992; Rittenhouse & O'Brien 1996).

Some, but not all ESDs, excluded patients with contraindications to the treatment. This has implications for the comparability of the no-treatment control group, as in studies without such exclusion criteria, one would expect to see a comparatively larger proportion of patients with contraindications in that group. Database designers should ensure that relevant variables for adjusting are available.

Very few studies allude to an intended target population (of which the sample may be representative), and sample characteristics are rarely compared with those of target populations or excluded patients. Whereas there was some indication that ESDs may have included older and more severely ill patients than RCTs (e.g. case study 2), Eggleston's inclusion criteria (case study 4) are more restrictive than those in comparable RCTs, most probably because the authors sought to avoid baseline differences in severity and confounding by indication.

In summary, it is disappointing to observe that claims for greater generalisability in ESDs were rarely substantiated in the Review. Even retrospective ESDs should be able to assess and report the extent to which their final sample is representative of a wider target patient population to which the authors wish to generalise their findings; as a minimum, authors need to comment on existing limitations for doing so, as was done in the Dornase Alfa Case Study.

## 8.4 Data protection

Given the difficulties, the Dornase Alfa Case Study experienced due to data protection considerations, and the importance of this subject for similar epidemiological research, this section is dedicated to relevant issue for ESDs, and puts these into the context of current UK guidance and ongoing debate.

The identified published ESDs showed a remarkable lack of consideration of confidentiality and data protection issues. Hardly any paper mentioned whether data were anonymised, and only three reports considered patient consent in any form. This may point at a then still relatively relaxed approach to the use of patient data for research. Given the recent changes in regulations concerning the use of patient data for research (and international variations in these), it is impossible to judge, whether studies would have fulfilled any legal requirements applicable at the time.

During the Dornase Alfa Case Study, the NHS Trusts' decision-making on the data protection and confidentiality questions were hampered by uncertainties surrounding the interpretation and application of new laws and guidelines coming into force at the time. In the NHS, these developments coincided with the introduction of a new system of "Caldicott Guardians" in every Trust. These guardians and their local teams and advisors were only beginning to learn to exercise their new responsibilities. Similarly, the role of research ethics committees in data protection and confidentiality was rather unclear at the time. In addition, the concurrent negative publicity of medical research misconduct contributed not only to the development of new procedures and guidelines, but also to the insecurities felt within hospital Trusts, particularly children's hospitals (Jones 2000; Woodman 2000).

Despite one Trust data protection officer initially deciding otherwise, others clearly interpreted the Data Protection Act 1998 as demanding explicit patient consent for an outside researcher accessing patients' medical records. This was at a time when well-meaning clinicians might still have felt in control of decisions about their patients' records and been inclined to grant a researcher or fellow health professional access to records for research based on trust. However, such disclosure would have constituted a clear breach of confidence (Romano-Critchley & Sommerville 1999); at the time, this was much less clearly understood by clinicians.

Today, the implications of the Data Protection Act 1998 and Health and Social Care Act 2001[6] are very widely appreciated, and application processes for access to patient data are increasingly streamlined. Only recently (November 2008), a new Coordinating System for gaining NHS Permissions (CSP) has been introduced to the NHS in England. Starting with National Institute for Health Research Clinical Research Network "portfolio studies", this will streamline the process of applying for permissions for new research and reduce duplication across NHS organisations. Through a single entry point via the Integrated Research Application System (IRAS), researchers will be able to apply for permissions from all NHS sites through one point. The application covers information relevant to approvals from a range of bodies, including research ethics committees, local NHS research offices, and the Patient Information Advisory Group (PIAG) (now subsumed into the National Information Governance Board). Arguably, such a

[6] Section 60 of the Health and Social Care Act 2001 permits processing of confidential patient information in specified circumstances where it is not currently practicable to satisfy the common law confidentiality obligations, e.g. seek consent.

The Act also allows for disclosure of information to specific bodies for specific purposes; these have to be approved by Parliament. The Health Service (Control of Patient Information) Regulations 2002 (Anonymous 2002) (for England and Wales only) were the first to be made under the Act. They support the operations of cancer registries in respect of medical purposes (including research) related to cancer, and the Public Health Laboratory Services (PHLS, now the Health Protection Agency) in respect of communicable diseases and other risks to public health (mainly in terms of control and monitoring them). Interestingly, research is not listed as one of the purposes for which confidential information may be processed by the PHLS.

streamlined approval process could prevent some of the discrepancies in local decision-making which we experienced with the Dornase Alfa Case Study. One may hope that this process reduces the time- and resource-intensive application and decision-making phases which reportedly have dogged epidemiological research for some time now (as illustrated by our own publication (Strobl *et al.* 2000) and others which followed, such as recently Metcalfe *et al.* (2008), who demonstrated that a PIAG application alone lost their team £560,000 of funded research time).

However, it is likely that the access to at least some disease databases and research databanks may have to be negotiated separately at least for some time. The process for applying for the use of the ERCF had not been devised prior to our application. But there is evidence that at least some database operators are already much clearer about their relevant access and governance processes, including information governance and intellectual property agreements (examples include the UK CF register at Dundee University, the MRC and Wellcome Trust report on access to collections of data and materials for health research (Lowrance 2006), as well as work under the NHS Connecting for Health Research Capability Programme (Care Record Development Board 2007)).

At a very early stage of the Dornase Alfa Case Study, an honorary contract was sought with one of the Trusts, in order to bring me as a researcher and health professional into the "NHS family" and thus within the sphere of influence of its disciplinary procedures. Again, this was not uncommon practice at the time. However, it would have been difficult to argue that I was involved in the care of patients whose records I would have accessed. Whereas audit is sometimes argued to be part of the healthcare process - and thus does not require consent - , the same cannot be said about

research.[7] Therefore, it was unclear, how a court of law might have looked upon an honorary contract in this context. The Department of Health now recommends the use of a "Research Passport" for researchers not employed by the NHS. This provides a mechanism for streamlining the pre-engagement checks for such researchers and assuring NHS organisations that they have taken place, and thus is effectively akin to an honorary contract with one organisation which is recognised by others. It must be stressed that this does not permit researchers access to data (without consent) they would not otherwise have had.

The uncertainty of how anonymous data was to be defined impinged on the confusion as to whether consent was required for the use of the ERCF in the Dornase Alfa Case Study. The Data Protection Commissioner held that any coded data was personal data, regardless of the location of the decoding key. This is a very contentious issue with epidemiologists, as much of their research would simply be impossible to conduct on totally anonymous data, because updating longitudinal data or validating any data would be impossible.

Understandably, there was much protest and ample misunderstanding around this issue in the literature, where some guidance documents clearly saw coded data as legitimate material for research, without patient consent. This certainly seems to have been the view in the BMA's guidance (Romano-Critchley & Sommerville 1999), albeit that it advised members to seek legal advice before using anonymised data. In many guidance documents it was not clear, whether the authors distinguished whether decoding keys still were still in existence or not.

[7] This leads to the question of whether audit involving personal data but performed by persons not involved in the care of patients concerned (e.g. audit assistants employed by or contracted by the Trust) requires patient consent. Some Trusts may routinely enter honorary contracts with such audit personnel in an attempt to resolve this issue. It would seem that there is at least a theoretical temptation to brand a project "audit" rather than "research", particularly where there is uncertainty about the distinction.

Thus, the Data Protection Commissioner was at odds with at least some of them. In the eyes of the Information Commissioner, the anonymised ERCF data constituted personal data to the Trusts. The act of anonymisation is seen as an instance of processing (Information Commissioner 2001). The Commissioner's letter implied that the data also remained personal data to the researchers. However, his more recent guidance opens the possibility for a different interpretation (Information Commissioner 2001): the Commissioner implies that the data controller must hold or be likely to come into possession of information which could identify individuals. She states *"whether or not data which have been stripped of all personal identifiers are personal data in the hands of a person to whom they are disclosed, will depend upon that person being in possession of, or likely to come into the possession of, other information* (Information Commissioner 2001, p.14). Some clarity seems to have emerged over time, and definitions appearing in more recent documents are increasingly congruent along the lines of the MRC definitions in Table 8.1. Thus, it can be argued that the data were not actually personal to myself, and current definitions of pseudonymised data would bear that out. Nevertheless, there is still room for interpretation as to where exactly the line lies between anonymised and pseudonymised data (Fistein 2008).

Similarly, it seemed impossible to extract a clear answer from guidance documents as to whether disclosure of anonymised data constituted a breach of confidentiality. However, if one accepted that anonymised data was personal data, then the disclosure of the data to the researcher as a third person would have had to conform to the Act. According to the Act, an individual's consent is needed for disclosure of their personal information. Helpfully, the House of Lords decided a case[8] recently which addressed a number of issues with a direct bearing on what constitutes personal information. It gives comfort to the view that:

*". . . pseudonymous information may be disclosed like anonymous information so long as the key to the re-identification is only held by the*

[8] Common Services Agency v Scottish Information Commissioner – 9 July 2008

*discloser. This may be of considerable significance to those in the health sector, who often need access to uniquely coded data for research purposes, but where the recipient of the information does not need access to the code."*

**Table 8.1: MRC Data & Tissues Tool Kit - Definitions of Anonymised Data[9]**

| **Anonymised data / information** |
|---|
| Anonymised data are data prepared from information from which the person to whom is relates cannot be identified. The term is used when referring to robustly pseudonymised / linked data or unlinked anonymised data. |
| **Pseudonymised data, also referred to as linked anonymised** |
| This is anonymous to the people who receive and hold it (e.g. a research team), but contains information or codes that would allow others (e.g. those responsible for the individual's care) to identify an individual from it. |
| **Unlinked anonymised data, or truly anonymised data** |
| This contains no information that could reasonably be used, by anyone, to identify the individual or study participant. |

The earlier Department of Health document "*Building the Information Core: Protecting and Using Confidential Patient Information*" laid down the strategy for making the NHS compliant with legal and ethical requirements (Department of Health 2001a). The document considers that "consent [to processing] is not required where information has been effectively anonymised" (p.1). Anonymisation here includes pseudonymisation. However, patients still need to be informed.

The response letter from the Data Protection Commissioner's office was not very clear on this point. The letter suggested that "pseudonymised" (i.e. coded) data could be processed in accordance with the Act, if a well-constructed security protocol prevented the breaking of the code by anyone outside the CF centres/NHS Trusts. Crucially, the letter suggested that informing patients of the likely uses of their data (including disclosure for research) and giving them an opportunity to object to such uses was sufficient to comply with the First Principle of the Act. Thus in the eyes of

[9] The MRC Data & Tissues Tool Kit is now an authoritative source of guidance which has been developed in partnership with key NHS and research organisations. http://www.dt-toolkit.ac.uk/home.cfm

the Commissioner the Act did not demand explicit written consent for the given project. It could also be implied that information about the possible use for research in general - and not necessarily a particular project - is sufficient information. However, as stated above, colleagues in one Trust interpreted this differently. The problems such differences in interpretation of the Act or related guidance could pose for data access for multi-centre research were drawn to the attention of the medical community (Strobl *et al.* 2000). Current practice is for data protection officers to approve a research protocol before its submission to a research ethics committee. Other data protection officers are expected to accept the decision of the first officer, but are legally required to make their own decision. Clearly, such decisions always require a degree of judgement, and it is likely that officers are risk-averse in their decisions, as they may not be trained nor encouraged to assess the risks involved in permitting access to coded records.

There is a current fierce debate on the issue of whether researchers should be able to access patient records without consent to identify potential research participants who could then be approached to consent to research. An earlier draft of the new NHS Constitution had included a sentence which would have ensured that this was possible. This sentence was removed in the final version (Department of Health 2009). Nevertheless, the presentation of the issue on the National Institute of Health Research (NIHR) website permits some reading between the lines which illustrates that this debate is still not resolved.[10]

---

[10] Speaking at a summit in 2008, Mr A Johnson (then Health Secretary) said:

> *"I want every patient in the NHS to have the right to take part in approved medical research that is appropriate for them, if they choose to...."*

[the NIHR web text continues:] This means that in future, all patients in the NHS will have the right to expect that their health record will be used... to identify whether they are suitable to take part in approved research which is relevant to them. Appropriate patients will be notified of opportunities to join in, and will be free to choose whether they wish to do so, after a full explanation. Speaking at the Health Research Summit, Harpal Kumar, Chief Executive of Cancer Research UK, said:

> *"The Government's announcement today is extremely welcome given that it helps to place health research at the core of the NHS. It shows a commitment to provide more information about the health research taking place in this country, and to involve and recruit more patients in clinical trials...."*

From some guidance documents it might be inferred that consent to the use of data for research may be implied where patients are informed that their data could be used for this purpose and do not object. A similar position was presented in the letter from the Data Protection Commissioner's Office. However, routine mechanisms to merely inform patients in advance about the potential use of their personal data for further research (for example, through notices in waiting rooms) may not be seen as constituting sufficient consent (personal communication with a data protection officer). It was also somewhat unclear, whether patients who do not register their refusal can be said to have consented. Essentially, I depended on the Trusts' interpretation of whether their (routine) patient information processes about possible uses of patient data were sufficient to comply with the Act. However, the nature of consent required is still not entirely clear today (Fistein 2008).

In 2003, the Department published its code of practice for NHS staff in England and Wales on the issue of confidentiality (Department of Health 2003). Explicit/express consent is defined as "*clear and voluntary indication of preference or choice, usually given orally or in writing and freely given in circumstances where the available options and the consequences have been made clear.*" (p.5) Such consent is needed for disclosure of identifiable data for research, unless public interest justifies the disclosure, or where the Health & Social Care Act 2001 Section 60 supports disclosure temporarily. It is made clear that patients' agreement with their use of data for research cannot be assumed, as this is not part of the healthcare process.

The Commissioner's recent documents (Information Commissioner 2001; Information Commissioner 2002) provide some better clarification on the nature of consent required than was available at the time the Dornase Alfa Case Study project evolved. The Commissioner considers that consent does not necessarily have to be in writing. However, a non-response to

communication or information cannot be considered consent. Explicit consent should be absolutely clear in terms of specific details of processing, including purposes and nature of processing and any specific aspects relevant to the individual. Thus to inform patients at ERCF registration specifically about the possibility of later disclosure of the anonymised ERCF data for particular types of research, and giving patients an opportunity to object to this (as well as documenting their response) could be argued to be sufficient to satisfy the requirements of the Act in relation to the Dornase Alfa Case Study. Whereas written consent is not required, it would seem prudent to seek such documentation of a patient's agreement.

The debate as to whether consent to the use of medical records in research should be required at all continues today with committed proponents on both sides of the argument, those favouring consent on the basis of patients' right to know and understand the use of their data (Clayton 2008), and those defending the feasibility of epidemiological research, including leaders of large research funding charities (see Cooper 2009). Each side claims support from patient and public views, but repeated calls for wider public debate have largely remained hollow, as the debate remains limited to research and legal-ethical fora and relevant NHS bodies.

Finally, the status of medical research itself within the Data Protection Act remains painfully uncertain, as Fistein (2008) explains in a working paper on the issue of consent for the NHS Connecting for Health Research Capability Programme. Fistein points out that the Act permits the use of health data for research as long as a "Schedule 2" condition is met. One of these conditions refers to processing necessary "in the public interest". Fistein argues that this point in particular (as well as other Schedule 2 conditions) needs urgent clarification in relation to research.

The problems are thus far from resolved, resulting in an ongoing situation which is well summarised in the still topical words of the Academy of Medical Sciences from 2006:

*"The legal framework around the use of personal data in research is a complicated patchwork involving UK legislation, case decisions and European directives, augmented by various guidance documents. There are many areas of imprecision, and the courts have not tested the legislation as it applies to medical research. ... The resulting variable legal interpretations have been a source of great difficulty, delay and disillusionment for researchers."* (Academy of Medical Sciences, 2006, p. 3)

## 8.5 Next steps for ESDs

Notwithstanding the slow progress in addressing data protection issues in ESDs and databases, and the un-exploited potential of greater generalisability of ESDs, the main remaining concerns centre on the internal validity of such studies. Earlier sections in this chapter have identified a considerable number of potential sources of bias and misclassification which were identified or might have been suspected in some of the studies included in the Review. Whereas prospective cohort studies have at their disposal methods for better ensuring validity and accuracy of the data, the mostly retrospective ESDs included in the Review were limited by the data they had available. Serious methodological criticisms have in the past been levied against disease registries (Lewis 2001), and researchers using them for ESD have limited options for retrospectively correcting any shortfalls. They can only carefully assess them - and the DQR has shown the practical problems in even doing that -, and adapt the study design accordingly, risking a loss of validity of the study. Any use of pre-existing data in ESDs therefore must be approached with extreme caution and should probably remain limited - if used at all - to assessments of large treatment effects, provided study validity is clearly established.

So what about prospective data collection? An ESD of prospectively collected data is methodologically a prospective cohort study (i.e. the outcome develops after study initiation), the qualities of which are described in any epidemiological textbook; such studies tend to be awarded a relatively higher position in hierarchies of research evidence. The crucial point, however, is that some defenders of database studies seem to argue for setting up databases designed without specific (effectiveness) hypotheses in mind (Rawlins 2008). This means that ESDs of such prospective databases may not benefit from bespoke data definition and

collection designed to control specific biases, which may well negate any benefits of prospective data collection. The methodological limitations of cohort studies compared to RCTs are not going to be addressed by such an approach; rather one set of limitations would be exchanged for another.

The goal for effectiveness research surely has to be to capitalise on the vastly increased amount of data potentially available thanks to information technological advances, without compromising validity. The Dornase Alfa Case Study is an excellent example for testing the requirements; it should have been as close as anything to a hypothesis-testing design (admittedly, and perhaps surprisingly, the stated aims did not imply that). Nevertheless, the methodological issues were considerable, and improvements suggested based on the work here need to be heeded in the design of hypothesis-specific databases.

This underlines the need for a more controlled approach to setting up databases where data collection methods and procedures as well as variable definitions are rigorously developed, piloted and their reliability and validity assessed before implementation. This might most meaningfully be achieved by focussing on a very small number of new databases. These could be set up to allow comparison with concurrent RCTs to enable more learning on the validity of ESDs. This experience should then inform the design of prospective ESDs in areas where RCTs are no longer possible (e.g. where equipoise is lost). This may affect many drugs in the post-marketing phase, and post-marketing surveillance databases and their use may be an appropriate focus for such methodological research.

Rawlins (2008) seems to favour setting up databases to create potential research resources for the future, but not necessarily with any specific hypothesis in mind (he strengthens his arguments by referring repeatedly to the use of databases in identifying adverse outcomes, rather than beneficial outcomes). He represents a wider current enthusiasm for collecting data in various clinical areas, often for rare conditions or

interventions, in the hope of being able to undertake future research - not infrequently on effectiveness and outcomes. This enthusiasm seems to span the HTA community, industry, as well as clinical communities. The ERCF is a stark reminder that even meticulously planned databases can encounter a whole host of unforeseen problems in the analysis and interpretation of subsequent effectiveness evaluations. It seems therefore unlikely that databases set up with no specific research focus and extremely tight controls of data quality and completeness in place could ever become reliable and worthwhile future resources for effectiveness research.

The lack of any systematic approach to assessing and discussing such potential sources of bias in study reports, however, remains a major concern, particularly since the Review was limited to high-quality journals. In the interest of safeguarding and promoting the quality of such reports, more systematic and rigorous approaches need to be applied to peer review processes. The British Medical Journal has recently suggested that authors should make raw data available to peer reviewers; in a similar way, peer reviewers of ESDs should routinely be furnished with all necessary evidence to judge the quality of the source database, its appropriate use, and the interpretation of findings by the ESD authors. The recently published guidance documents represent useful tools for this (Huston and Naylor 1996; Motheral and Fairman 1997; Gliklich & Dreyer 2007; van Elm *et al.* 2007; Berger *et al.* 2008).

Once study quality of ESDs has been established, the interpretation of their evidence in relation to a body of evidence requires further consideration also. The comparison case studies have highlighted that direct comparison with RCTs may be problematic for several reasons, and the results of ESDs may be misinterpreted or misused. It may be argued that a direct comparison is inappropriate as both study designs may (and arguably should) address different hypotheses. The literature on synthesis methods suggests systematic approaches to assessing rigorously the contribution of

ESDs to any evidence base (GAO 1992; Labin 2007). The potential of these methods specifically in drug effectiveness evaluations warrants further development.

# 9 Conclusion

## 9.1 Summary of Findings

I had set out to describe and explore the methodological challenges of ESDs by undertaking a drug effectiveness case study using an existing database (the Dornase Alfa Case Study), comprehensively reviewing comparable published studies, and by comparing the findings of ESDs with RCTs of the same research question. The path has not been a smooth one! The Dornase Alfa Case Study highlighted considerable problems with undertaking effectiveness studies on pre-existing databases, particularly around data quality and completeness, but also regulatory issues on access to health data for research. In turn, the review of published ESDs was limited by the poor bibliographic indexing and reporting practices of such studies, and the four comparison case studies were limited by their diverse nature. Nevertheless, the thesis' resulting contribution to knowledge is considerable and centres on a number of achievements arising from the Dornase Alfa Case Study, Review, and comparison case studies with RCTs. These are summarised in turn:

Firstly, the Dornase Alfa Case Study - despite not having fulfilled its original aim of evaluating long-term effectiveness of the drug - has identified important methodological lessons through the detailed approach taken. These have been published in high-ranking peer-reviewed journals (Strobl *et al.* 2000; Strobl *et al.* 2003):

The data quality and DQR processes included the first direct validation of the ERCF database and demonstrated that some key outcome variables were poorly defined (e.g. exacerbations) and incompletely reported (e.g. exacerbations, deaths). Also other variables lacked clear definitions and comprehensive reporting. The work demonstrated that despite extensive routine data verification processes built into the ERCF operation, relatively trivial data checks as part of data cleaning here identified further errors in

the key variables. The ERCF had also ignored important softer information and centre differences in clinical and reporting practice; these would not have been available to analysts who did not know the situation in the centres in sufficient detail. There was a lack of control of patient transfer processes, resulting in discontinuity of records. I identified possible sources of bias arising from the data collection processes (e.g. information bias resulting from different frequencies of follow-up visits and reports). These findings represent lessons for databases in general. Nevertheless, the recent seminal work on disease registries by the Agency for Healthcare Research and Quality (Gliklich & Dreyer 2007) presents the corresponding US sister database of the ERCF, the ESCF, as one of its case examples. Thus even highly regarded and intensively resourced registries may show considerable flaws when carefully examined.

The Dornase Alfa Case Study also gave rise to a detailed examination of the uncertain legal and guidance framework relating to confidentiality and data protection applicable to research using health records. At the time, there was little clarity and plenty of conflicting guidance and advice. I explored the existing law and guidance in an attempt to answer the relevant questions of consent and access to (coded) patient data for the Dornase Alfa Case Study, and ultimately published these vital concerns for the epidemiological research community in a paper in the British Medical Journal. This is still a much-debated area of law and ethics today.

The Dornase Alfa Case Study identified that generalisability - an oft-quoted potential strength of observational studies - is difficult to assess and even more difficult to achieve in databases. This is because they depend entirely on the registration of individual cases, usually through clinical centres which may not be representative of the relevant target patient population of interest. The extent of patient recruitment to the ERCF varied between participating centres, and clinicians considered that shared-care patients (who receive only part of their care from a specialist centre) should have been distinguished from the main cohort. Database

operators (and analysts) may thus remain in the dark about the extent of completeness of the registration of all patients from each centre, as well as the representativeness of the sample vis-à-vis the target population of all CF patients.

Lastly, the Dornase Alfa Case Study provides descriptive data for comparison with other published data on CF populations, and it uses existing recent observational studies on the effectiveness of dornase alfa to discuss the merits and pitfalls of multi-variate analyses in what remains a challenging patient population for epidemiological research.

Secondly, the findings of the comprehensive review of ESDs of drug therapies presented here represent a further major achievement. Much has been written on databases and also their use in HTA (Lewsey *et al.* 2000; Williams *et al.* 2003; Raftery *et al.* 2005; Gliklich and Dreyer 2007), and on how database studies should be reported (Motheral and Fariman 1997; von Elm *et al.* 2007), but so far no systematic review of studies having used databases for (drug) effectiveness research has been reported.

The Review identified that increasing numbers of ESDs are being published. It is likely that a significant publication bias is at play here, as by far the majority of identified studies reported a significant effectiveness outcome. Most studies were retrospective, and perhaps surprisingly few reported direct or indirect industry funding. This suggests that this relatively in-expensive research design is more broadly accessible to other clinical and research communities. Nevertheless, there is scope for significant industry involvement (and control), particularly through post-marketing surveillance databases; this could limit - and at worst distort - the use of such resources for genuine HTA.

Issues of consent and data protection appeared to concern a worryingly small minority of authors; this was probably a sign of the time and may be explained by data protection regulations having come into force only

relatively recently. An equally surprisingly small number of studies made any attempt to demonstrate their representativeness of a target population or their greater generalisability compared to randomised trials. This is disappointing, given the theoretical potential of ESDs in this area.

Many authors did not report any assessments of data quality, which raises doubts as to whether rigorous data quality assessments such as the ones applied in the Dornase Alfa Case Study are the rule in published studies. The definition of exposure variables (drug treatment) was often sufficiently clear, but the outcome variables of some studies relied on proxy measures (e.g. prescribing). Very few studies were able to refer to data validation against external information sources. Most studies attempted to adjust for confounding in the analysis, however, the methods used were simple, and the success of any statistical adjustment was rarely reported. In summary, the overview of published ESDs provided by the Review does not give any cause for confidence in the current use of this study design.

Thirdly, the four case study comparisons of ESDs with available RCTs having addressed the same effectiveness question, add to the knowledge available from other reviews which did not explicitly focus on ESDs (e.g. Britton *et al.* 1998; Benson and Hartz 2000; Ioannidis *et al.* 2001; Kunz and Oxman 1998; MacLehose *et al.* 2000). There was no detectable systematic pattern of differences in results between study types. However, the true comparability of ESDs and RCTs (in terms of patient selection, or outcome variables) could be questioned, and this may explain some discrepancies in results. Nevertheless, this may also apply to some of the previously published reviews.

In two case study comparisons (1 and 3), the results appeared to concur, but the comparison could not be quantified, and in one case may be entirely spurious due to variations in outcome measures (case study 3). In the remaining two case studies (2 and 4) the ESDs claimed equivalence of effectiveness of two comparator treatments, whereas the RCTs favoured

one particular treatment. Both ESDs had some industry involvement, and in one case (case study 4) the findings were later mis-represented in a cost-effectiveness analysis in favour of a drug which is nowadays clearly not the drug of first choice (but manufactured by the drug company leading on the ESD). In the second case (case study 2) the difference in results may be partly explained by patient selection and insufficient follow-up in the ESD - decisions which might have been made without prejudice.

In terms of frequently cited benefits of ESDs vis-à-vis RCTs, ESDs were often larger than comparator RCTs, but not as a rule; similarly their follow-up was not always significantly longer than RCTs', and in case study 2 most likely too short. Equally disappointing is the lack of emphasis of ESDs on demonstrating their generalisability, rather than just claiming it and using it as a justification for the chosen research design. Exclusion criteria of ESDs often mimic those of RCTs, but pragmatic ones often have to be added, particularly due to poor data quality or missing data; this itself can introduce selection bias and jeopardise generalisability.

Thus, the four case study comparisons have highlighted some strengths of ESDs vis-à-vis RCTs (e.g. the ability to measure outcomes over a long time frame and large populations), but probably more pitfalls. In addition, the poor reporting and seemingly lack of quality control of this design in peer review processes, as well as the likely control of many large databases by industry pose considerable potential for biased use.

## 9.2 Conclusions

There is considerable interest in ESDs in the HTA community, as demonstrated by a recent paper by Rawlins (2008), the Chair of the UK's National Institute for Health and Clinical Excellence, in which he underlines the virtues of observational research. There also appears to be a degree of consensus in the recent literature on the complementary values of randomised and observational study designs and the fact that neither are immune to bias or provide perfect information (Rawlins 2008; Sørensen *et al.* 2006). Albeit that at times such discussions - although acknowledging the value of both designs - are largely limited to pointing out the weaknesses of one study design, and arguing for the other on the basis of emphasising its strengths.

The rising importance of ESDs is further underlined by the increase over time in the numbers of published ESDs identified by the Review, and the recent publications of relevant guidance on disease registries as well as observational research (Gliklich & Dreyer 2007; von Elm et al. 2007). Information technological advances in healthcare have vastly increased the potential for routine health data collection and the theoretical potential for ESDs.

Clinical communities as well as health technology assessors often show considerable enthusiasm for database projects to capitalise on the increasing computerisation of clinical systems. However, little consideration seems to be given to the later uses of such databases. Rawlins (2008) argued in favour of observational studies, largely based on their known strengths in investigating adverse effects of treatments. It would be entirely inappropriate to conclude that such designs are equally well suited to investigating beneficial effects of treatments (i.e. effectiveness). There are after all many examples of observational study

findings having misled the clinical community (and thus harmed patients) until robust trial and systematic review evidence emerged (MacMahon and Collins 2001). Rawlins is right in arguing for the need for judgment in interpreting evidence, but the first question in judging an evidence base is still: is the evidence reliable? Current ESD publications do not permit a confident positive response to that question.

The overwhelming conclusion of the thesis is that despite the current enthusiasm for databases and disease registries, and the implicit hope that they could successfully be used for future ESDs, published ESDs are poorly reported, and the methodological challenges are poorly appreciated and addressed. Despite theoretical awareness of methodological solutions, existing examples leave much to be desired. Through meticulous detailed and critical analysis, this thesis provides first-hand evidence of actual analytical and practical problems associated with the retrospective use of databases and highlights severe limitations.

The ERCF might have been seen as one of the most promising disease databases for an ESD, given that its primary focus was on one drug treatment, it involved control patients, and above all, it operated intensive routine data quality assurance processes. Nevertheless, the Dornase Alfa Case Study essentially was unable to undertake a reliable effectiveness analysis on the available sample, and no credible effectiveness analysis has ever been achieved and published from the complete ERCF dataset either.

The findings from the presented Review of other published drug ESDs did little to improve confidence in the current quality of such studies. Together with the Dornase Alfa Case Study, the Review demonstrated that a variety of identified data quality problems, as well as confounding by indication, are very difficult to address within a retrospective study design and thus result in further study limitations (e.g. additional exclusion criteria having to be applied, which may limit generalisability). In addition, there is a strong indication that publication bias is at play here,

and the poor quality controls seemingly applied by peer reviewers do little to limit potential bias arising from commercial interests.

Glicklich & Dreyer (2007) provide detailed guidance on the development and use of registries, and example databases or ESDs quoted by them are presented for illustrative purposes, but not critically reviewed. The Review presented here demonstrates that the application of existing methodological knowledge is still limited in this area. The contribution of existing published ESDs to an evidence base has to be severely questioned not merely on the basis of methodological weaknesses, but simply on the basis of poor reporting which does not allow any comprehensive appraisal of the validity and generalisability of many such studies.

The findings of this thesis draw attention to a considerable number of potential and actual shortfalls of published ESDs, which are largely retrospective, and their value in effectiveness research hence has to be questionned. It is clear that observational studies cannot replace RCTs, particularly where effects are moderate or small (MacMahon and Collins 2001), and neither do RCTs make observational studies redundant. Nevertheless, ESDs have undeniable strengths, such as potentially better generalisability, larger sample sizes, and longer follow-up periods under naturalistic conditions, but these are currently not being exploited. Also, at least theoretically, ESDs should be able to fulfil key Bradford-Hill criteria of causality[11] (Hill 1965) and have the potential to contribute much to an effectiveness evidence base; and there will always be scenarios where RCTs are simply not possible. However, most existing drug ESDs are not of sufficient quality to be relied upon as sole sources of effectiveness evidence. Thus the current question is not necessarily whether ESDs could or should be used where RCTs are not possible, but how we can improve their quality and validity so they add value to effectiveness evidence.

[11] Strength and association, consistency, specificity, temporality, dose-response, biological plausibility, biological coherence, experimental evidence, analogy.

The answer to the question of whether ESDs are worthwhile pursuing is therefore: yes, if data quality and better control of biases are achievable. There are essentially two main ways forward. Firstly, the rigorous implementation of existing guidance as well as recommendations arising from this thesis. Together they constitute a considerable body of knowledge, which has probably rarely if ever been put into practice in existing databases; certainly the reviewed ESDs (albeit that they pre-dated some key guidance) do not demonstrate the application of much of this knowledge. The use of databases set up prospectively with a single planned hypothesis for an ESD is a more promising approach than the retrospective analysis of databases collected for different purposes.

In addition, internal and external quality control efforts need to be strengthened considerably. The latter should include external peer review and assessment not just of ESDs, but of databases themselves, regardless of ownership. Peer-reviews of ESDs submitted for publication need to apply more rigorously the available methodological evidence; a tool for study quality assessment would be helpful for this, and should be developed.

Secondly, more methodological research should support the quality improvement of ESDs. Research needs to address the role and implications of patient and clinician preference, and means for adjustment in data analysis. Such research will probably need to use comparisons between study designs in order to validate new approaches in ESDs. Methodological work could initially be limited to relatively uncomplicated settings (e.g. short-term treatment) to enable more focussed attention on particular features of ESDs at any one time (e.g. generalisability, control of confounding by indication).

Better quality ESDs would also enable further development and a more confident use of some promising approaches to evidence generation, particularly the randomised database study (Sacristán et al. 1998; Mosis 2006), and cross design synthesis (GAO 2002). The use of both has hitherto

been regrettably limited, and arguably they would themselves benefit from higher quality databases and ESDs being available.

Finally, the contribution of existing published ESDs to an evidence base has to be questioned not merely on the basis of methodological weaknesses, but also simply on the basis of poor reporting which does not allow any comprehensive appraisal of the validity and generalisability of many such studies. The review found that publication bias of ESDs is likely to be considerable, and that commercial interests may exploit the current lack of rigour and of widely accepted quality standards of ESD publications. To address these has to be one of the most urgent priorities in the short-term.

## 9.3 Recommendations

Whether ESDs are used within or outside methodological research, they ought to heed existing guidance as well as the recommendations arising from the work presented here. Recently the US Agency for Healthcare Research and Quality has published a seminal piece of guidance on the establishment and use of registries for evaluating patient outcomes (Gliklich & Dreyer 2007). Whereas this publication addresses many points identified in this thesis, most of the recommendations listed below are additional to those in the publication.

### 9.3.1 Recommendations for databases supporting ESDs

1. The thesis found that key variables were often poorly defined or absent and thus not usable for ESD analysis. In order to support ESDs, databases need to provide variables which permit an unambiguous description of baseline characteristics including co-morbidities and possible confounders, treatment decisions, exposure to treatment, and key outcomes.
    a. In particular, the rationale for why patients are commenced on treatment needs to be recorded; this will also lead to the identification of potential confounding variables to be controlled in the analysis.
    b. Prior use of drugs of interest (including prior to enrolment on the database) needs to be recorded for each patient so that treatment naïvity can be determined.
    c. In order to characterise the exposure to the treatment of interest, databases need to include reliable assessments of compliance with treatment.

d. Variable definitions, as well as data collection processes, need to be robustly defined and piloted with participating clinicians / centres.

2. Far more attention than hitherto needs to be paid to data quality and completeness. It is unlikely that databases which collect dozens of variables can realistically dedicate sufficient efforts to ensuring the quality of all of them.
    a. A prudently selected small number of variables should be collected. This needs to be informed by the anticipated analysis. Collecting any data possible "for a rainy day", i.e. unclear research hypotheses, is likely to risk significant data quality problems.
    b. Routine validation against external sources of information (e.g. original records or other information sources) has to ensure the data quality of at least key variables, including exposure and outcome variables.
    c. Cross-checking of variables against each other (e.g. age and height) and also checks of consistency over time (e.g. of body weight or height) need to be part of the data quality routine.
    d. Information on treatment timing and duration enables better categorisation of exposure to the treatment of interest, particularly over a longer follow-up period. Timing errors (particularly where the year is wrong, e.g. where a course of antibiotics is recorded to have lasted for one year AND two weeks, when normal practice is a 2-week course) should therefore be routinely assessed.
    e. Imbalances in reporting and likely under-reporting (e.g. from different centres or on particular variables) need to be identified and investigated routinely.
    f. The Dornase Alfa Case Study identified differences in interpretations of variable definitions between centres. Contributing clinicians and centres need to be bound to consistent unambiguous variable definitions and reporting

practices; incentives for participation should be linked to data quality and consistency in reporting, rather than numbers of patients registered.

g. Sharing the intended investigative hypothesis with participating clinicians and centres will improve their cooperation, but will also be a source of assessment bias. Outcome assessors should therefore be blinded if at all possible, unless outcomes are unambiguous (e.g. death).

h. The Dornase Alfa Case Study identified a loss of continuity of records for patients transferred between participating centres. Clear processes for transfers of patients between centres need to be established to avoid discontinuity of records.

i. Database operators need to remain close to clinicians to understand softer issues relevant to the analysis, as well as changing practice over time.

3. Database operators need to consider how their database can demonstrate and verify the representativeness of the registered patient population against the source population.
4. Databases need to ensure they collect data in a way that is conform with current data protection legislation. This means that consent of patients for inclusion on the database but also for subsequent studies based on the database should be sought before patients are registered.
5. Database operators (and funders) need to define conditions and processes for access to and use of the data they hold. Given the possibility of industry sponsors limiting the use of database resources for bona fide research, patients should be made aware of any proposed access restrictions for such research prior to giving their consent.
6. Clear rules should be specified by regulatory authorities for the setting up and operating of post-marketing surveillance databases

and their use (and access to their data) to ensure they are available for research.

### 9.3.2 Recommendations for the conduct and reporting of ESDs

These recommendations have arisen from the work on this thesis and complement those identified in previous and subsequent guidance (Berger *et al.* 2008; Huston and Naylor 1996; Motheral and Fairman 1997; van Elm *et al.* 2007). An over-arching recommendation therefore is to amalgamate these into a widely-publicised consensus document. Much of the criticisms this thesis levelled against published ESDs might have been avoidable if researchers had followed such guidance in conducting and reporting ESDs.

1. Database analysts need to demonstrate clearly why an ESD should be undertaken rather than an RCT, including how the ESD hypothesis is different from one which a potential randomised study could address. If an RCT is impossible or unnecessary, the reasons need to be reported.
2. Database analysts need to ensure themselves of the quality of the database (see above recommendations). It is unlikely that non-condition-specific databases such as Mediplus or claims databases, and maybe GPRD to a lesser extent, will be able to meet these recommendations. They will remain problematic for much effectiveness research, particularly of drug treatments, as the indication for treatment is often not recorded, outcome measures might need to rely on proxy measures, and determination of exposure will be subject to errors and potentially bias.
3. Many published studies do not address the strengths of ESDs or only pay lip service to them; ESDs need to capitalise on their strengths; part of this is the demonstration of true greater representativeness and external validity.

a. The implications of and reasons for inclusion and exclusion criteria (both of the database as well as the ESD) need to be considered and reported.
b. Studies need to report the number of patients screened for inclusion into the study, and to assess and report the characteristics of excluded patients compared to included patients.
c. Comparisons of the sample with potential target populations should be undertaken and reported to permit judgments of likely representativeness.

4. Analysts need to familiarise themselves with recording and clinical practices within contributing centres to be able to identify and interpret systematic differences in the data. Data cleaning needs to focus on centre differences in reporting frequency, systematic differences in values of variables, and variations in outcomes.
5. The Dornase Alfa Case Study found that centre differences played a considerable but hidden role, but were rarely considered by published ESDs. Such centre differences need to be taken account of in any analysis.
6. ESDs need to consider the use of newer analytical methods better suited to controlling confounding by indication (see Berger *et al.* 2008). Unadjusted and adjusted results should be reported, and the likely extent of residual confounding explored.
7. Treatment decisions need to be taken into account in the analysis.
8. A sensitivity analysis should explore the likely impact of potential errors in the data.
9. The potential for under-reporting of variables needs to be assessed carefully and reported, and in particular, any possible non-randomness of under-reporting.
10. ESDs should explicitly address each potential source of bias, particularly selection bias, information bias, measurement bias, differential misclassification.

11. ESD reports need to record the possible variables and how the choice was made for using particular ones, and how these were defined.
12. Similarly, the quality assessments of individual variables should be reported.
13. ESD analysts need to consider the implications of lengthy follow-up in terms of changes in clinical practice.
14. Intention to treat analyses should be undertaken as a preference.
15. On the question of positive evidence for a causal relationship measured on the chronological relationship between cause and effect, the strength of a relationship, a dose-response relationship, consistency, and specificity (Elwood 1998), ESDs offer considerable potential. ESD reports should examine how they meet such criteria for causality.

### 9.3.3 Recommendations for further long-term effectiveness evaluations of dornase alfa

The ERCF was discontinued several years ago, and it is unlikely that the bulk of its data can be sufficiently well verified and improved to enable rigorous analysis of the long-term effects of dornase alfa. The most promising way forward to address any such hypothesis is for the existing UK CF database and other international database teams to join in a bespoke and pre-planned ESD, taking account of the lessons from the ERCF and of this thesis.

## 9.4 Suggestions for further research

- Whereas much has been written about the potential and potential pitfalls of observational studies as well as registries, the specific case of ESDs has barely been considered in the methodological literature. The increasing popularity of such studies needs to be accompanied by carefully planned research into their ability to make a valid contribution to an emerging effectiveness evidence base for treatments in routine use. This research should focus on the issues raised in this thesis and by others, and their actual impact on study validity. Post-marketing surveillance databases could provide a useful focus for this research.

- Methodological research on prospective ESDs should first focus on areas where comparisons between different designs are possible.

- Given the increasing use of ESDs, an instrument to assess the quality of such studies needs to be developed with some urgency. In addition to assessing dimensions relevant to observational studies, such an instrument needs to address the nature of the data source(s) and their use by a particular study, as well as focus on the potential biases of ESDs.

- Despite much rhetoric, the complementary strengths of study designs have not been fully exploited for effectiveness research. Cross design synthesis deserves further methodological attention.

- The publication bias of ESDs is worth examining, particularly in the area of drug research, where pharmaceutical companies may influence the decision to publish.

# References


Academy of Medical Sciences (2006) Personal Data for Public Good: Using Health Information in Medical Research. Academy of Medical Sciences, London.

Anonymous (1994). The Epidemiologic Registry of Cystic Fibrosis - Registry Reference Binder, unpublished.

Anonymous (1999). R v Department of Health, Ex Parte Source Informatics Ltd and (1) Association of the British Pharmaceutical Industry (2) General Medical Council (3) Medical Research Council (4) National Pharmaceutical Association Ltd (Interveners). Times Law Report. 14 June.

Anonymous (2000a). The Data Protection (Designated Codes of Practice) (No. 2) Order 2000. The Stationery Office Limited.

Anonymous (2000b). The Data Protection (Subject Access Modification) (Health) Order 2000. The Stationery Office Limited.

Anonymous (2000c). "Observational studies and randomized trials (Correspondence)." New England Journal of Medicine **343**: 1194-1197.

Anonymous (2000d). R v Department of Health, Ex Parte Source Informatics Ltd and (1) Association of the British Pharmaceutical Industry (2) General Medical Council (3) Medical Research Council (4) National Pharmaceutical Association Ltd (Interveners). Times Law Report. 18 January.

Anonymous (2002). The Health Service (Control of Patient Information) Regulations 2002. Statutory Instrument 2002 No. 1438.

Abbott, J., M. Dodd, J. Moorcroft, et al. (1995). Survival and lung function over a decade in a regional adult cystic fibrosis unit. 20th meeting of the European Working Group for Cystic Fibrosis, Brussels, Belgium.

Aitken, M. L., W. Burke, G. McDonald, et al. (1992). "Recombinant human DNase inhalation in normal subjects and patients with cystic fibrosis. A phase 1 study." Journal of the American Medical Association **267**: 1947-1951.

Alpert, J. (2000). "Are data from clinical registries of any value?" European Heart Journal **21**: 1399-1401.

Amarales, L., I. Dab, A. Malfroot, et al. (1998). Lung function studies before and after 12 months of DNase therapy. European CF Society Meeting, Berlin, Germany.

Amelina, E. L., A. Cherniak, N. Voznesenskiy, et al. (1999). "The effect of daily inhaled DNase in adult patients with cystic fibrosis." The Netherlands Journal of Medicine **54**: S81.

Antczak-Bouckoms, A., E. Burdick, S. Klawansky, et al. (1991a). "Using medical registries and data sets for technology assessment." International Journal of Technology Assessment in Health Care **7**: 123-128.

Antczak-Bouckoms, A., J. Tulloch and S. Klawansky (1991b). "Technology Assessment in the Normative Aging Study." International Journal of Technology Assessment in Health Care **7**: 143-149.

Badet, F., G. Bellis, M. deBraeckeler, et al. (2001). "Long survivants" in cystic fibrosis French population. 24th European CF Conference, Vienna, Austria.

Ballmann, M. and H. v. d. Hardt (2002). "Hypertonic saline and recombinant human DNase: a randomised cross-over pilot study in patients with cystic fibrosis." Journal of Cystic Fibrosis **1**: 35-37.

Bandolier (2002) "Treatment effectiveness and costs in reflux disease." Bandolier Library. www.bandolier.co.uk (accessed on 17 December 2002).

Barrett, G., J. Cassell, J. Peacock, et al. (2006). "National survey of British public's views on use of identifiable medical data by the National Cancer Registry" British Medical Journal **332**:1068-1072.

Bartošová, J., L. Smolíková, K. Kotíková, et al. (2000). Adherence to treatment in Czech CF patients receiving rhDNase. XIIIth International Cystic Fibrosis Congress, Stockholm, Sweden.

Bartošová, J., D. Zemková, P. Pohunek, et al. (1999). "RhDNase in Czech CF patients." The Netherlands Journal of Medicine **54**: S81-82.

Bate, C., J. Green, A. Axon, et al. (1997). "Omeprazole is more effective than cimetidine for the relief of all grades of gastro-oesophageal reflux disease-associated heartburn, irrespective of the presence or absence of endoscopic oesophagitis." Alimentary Pharmacological Therapy **11**: 755-763.

Bate, C., S. Griffin, P. Keeling, et al. (1996). "Reflux symptom relief with omeprazole in patients without unequivocal oesophagitis." Alimentary Pharmacological Therapy **10**: 547-555.

Benson, K. and A. Hartz (2000). "A comparison of observational studies and randomized controlled trials." New England Journal of Medicine **342**: 1878-1886.

Benvegnu, L., L. Chemello, F. Noventa, et al. (1998). "Retrospective analysis of the effect of interferon therapy on the clinical outcome of patients with viral cirrhosis." Cancer **83**: 901-909.

Berger, M. (2008). Design and Analysis of Non-Randomized Studies of Treatment Effects Using Secondary Databases. Report of the ISPOR Task Force on Good Research Practices for Retrospective Database Studies.

Bergmann, J., O. Chassany, J. Gandiol, et al. (1994). "A randomized clinical trial of the effect of informed consent on the analgesic activity of placebo an naproxen in cancer pain." Clinical Trials and Meta-Analyses **29**: 41-47.

Berlin, J. and G. Colditz (1999). "The role of meta-analysis in the regulatory process for foods, drugs, and devices." Journal of the American Medical Association **281**: 830-834.

Berlowitz, D., G. Brandeis and M. Moskowitz (1997). "Using administrative databases to evaluate long-term care." Journal of the American Geriatrics Society **45**: 618-623.

Berman, S., P. J. Byrns, J. Bondy, et al. (1997). "Otitis media-related antibiotic prescribing patterns, outcomes, and expenditures in a pediatric medicaid population." Pediatrics **100**: 585-592.

Bertolo-Houriez, E., I. Jacques, J. Derelle, et al. (1997). Effects of aerosolized recombinant human deoxyribonuclease (rhDNase) in cystic fibrosis. European Working Group for Cystic Fibrosis Meeting, Davos, Switzerland.

Bilton, D., L. Sharples, S. Elborn, et al. (1998). How frequent are pulmonary exacerbations in cystic fibrosis (CF) adult patients? European CF Society Meeting, Berlin, Germany.

Black, N. (1996). "Why we need observational studies to evaluate the effectiveness of health care." British Medical Journal **312**: 1215-1218.

Blais, R. (1991). "Using administrative data bases for technology assessment in health care." International Journal of Technology Assessment in Health Care **7**: 203-208.

Blau, H., Y. Yahav, G. Ginsberg, et al. (1997). Variable response to rhDNase therapy in cystic fibrosis patients with mild to moderate lung disease. European Working Group for Cystic Fibrosis Meeting, Davos, Switzerland.

BMA (2000) "The Impact of the Human Rights Act 1998 on Medical Decision Making." British Medical Association. http://www.bma.org.uk/ap.nsf/Content/the+impact+of+the+Human+Rights+Act+1998+on+medical+decision+makingm (accessed on 23 June 2003).

BMA (2003) "BMA Response to NHSIA & DOH National Confidentiality Consultation." British Medical Association. http://www.bma.org.uk/ap.nsf/Content/confidentialitycons (accessed on 23 June 2003).

Bowling, A. (1997). Research Methods in Health: Investigating Health and Health Services. Buckingham, Open University Press.

Bowman, L., D. Faulkner and L. Lucy (2000). "Differential costs and outcomes among pediatric oral antibiotic formulations identified from a large prescription database." Journal of Chemotherapy **12**: 72-78.

Bradley, J. M., L. Quinn and J. S. Elborn (2001). DNase therapy in CF: a survey of current use in the UK and Ireland. 24th European CF Conference, Vienna, Austria.

Braggion, C., D. Marchi, C. Castellani, et al. (1997). FEV1 decline between 7 and 12 years in relation to cystic fibrosis (CF) genotype: a cohort study. European Working Group for Cystic Fibrosis Meeting, Davos, Switzerland.

Brewin, C. and C. Bradley (1989). "Patient preferences and randomised clinical trials." British Medical Journal **299**: 313-315.

British National Formulary (2006). British National Formulary. London, British Medical Association and Royal Pharmaceutical Society of Great Britain.

Britton, A., M. McKee, N. Black, et al. (1998). "Choosing between randomised and nonrandomised studies: a systematic review." Health Technology Assessment **2**(13).

Brown, N., M. Melville, D. Gray, et al. (1999). "Relevance of clinical trial results in myocardial infarction to medical practice: comparison of four year outcome in participants of a thrombolytic trial, patients receiving routine thrombolysis, and those deemed ineligible for thrombolysis." Heart **81**: 598-602.

Buehler, J. (1998). "Surveillance". In: K. Rothman and S. Greenland: Modern Epidemiology. Philadelphia, Lippincott Raven Publishers: 435-457.

Butterworth, J., R. James, R. C. Prielipp, et al. (1998). "Do shorter-acting neuromuscular blocking drugs or opoids associate with reduced intensive care unit or hospital lengths of stay after coronary artery bypass grafting? CABG Clinical Benchmarking Data Base Participants." Anesthesiology **88**: 1437-1446.

Byar, D. (1980). "Why data bases should not replace randomized clinical trials." Biometrics **36**: 337-342.

Byar, D. (1991). "Problems with using observational databases to compare treatments." Statistics in Medicine **10**: 663-666.

Caldicott Committee (1997). Report on the review of patient-identifiable information. London, Department of Health.

Canto, J., C. Kiefe, O. Williams, et al. (1999). "Comparison of outcomes research with clinical trials using preexisting data." The American Journal of Cardiology **84**: 923-927.

Care Record Development Board (2007). Report of the Care Record Development Board Working Group on the Secondary Uses of Patient Information. http://www.connectingforhealth.nhs.uk/crdb/workstreams/secusesreport.pdf (accessed 27 March 09)

Carlsson, R., J. Dent, R. Watts, et al. (1998). "Gastro-oesophageal reflux disease in primary care: an international study of different treatment strategies with omeprazole." European Journal of Gastroenterology and Hepatology **10**: 119-124.

Carné, X. and J.-A. Arnaiz (2000). "Methodological and political issues in clinical pharmacology research by the year 2000." European Journal of Clinical Pharmacology **55**: 781-785.

Carter, S. (2002) "A Guide to the UK Legal System - Updated." http://www.llrx.com/features/uk2.htm (accessed on 25 June 2003).

Chaitman, B., T. Ryan, R. Kronmal, et al. (1990). "Coronary Artery Surgery Study (CASS): comparability of 10 year survival in randomized and randomizable patients." Journal of the American College of Cardiology **16**: 1071-1078.

Chew, D. P., D. L. Bhatt, M. A. Robbins, et al. (2001). "Effect of clopidogrel added to aspirin before percutaneous coronary intervention on the risk associated with C-reactive protein." The American Journal of Cardiology **88**: 672-674.

Chiba, N., C. De Gara, J. Wilkinson, et al. (1997). "Speed of healing and symptom relief in grade II to IV gastroesophageal reflux disease: a meta-analysis." Gastroenterology **112**: 1798-1810.

Choi, H. K., M. A. Hernan, J. D. Seeger, et al. (2002). "Methotrexate and mortality in patients with rheumatoid arthritis: a prospective study." The Lancet **359**: 1173-1177.

Christopher, F., D. Chase and R. Milne (1999). DNase for cystic fibrosis patients with mild to moderate lung disease. Southampton, Wessex Institute of Public Health Medicine, Development & Evaluation Committee Report No. 95.

Cimino, G., S. Diamanti, F. Alatri, et al. (1997). Efficiency of recombinant human DNase therapy in CF patients. European Working Group for Cystic Fibrosis Meeting, Davos, Switzerland.

Clayton, H. (2008). WResearch and the issue of patients' consent." The Lancet **372**: 2109.

Cobos, N., I. Danes, S. Gartner, et al. (2000). "DNase use in the daily care of cystic fibrosis: who benefits from it and to what extent? Results of a cohort study of 199 patients in 13 centres. DNase National Study Group." European Journal of Pediatrics **159**: 176-181.

Cochrane Collaboration (2001). Cochrane Reviewers' Handbook. Oxford, Cochrane Collaboration.

Collins, R. and S. MacMahon (2001). "Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials." The Lancet **357**: 373-380.

Concato, J., N. Shah and R. Horwitz (2000). "Randomized, controlled trials, observational studies, and the hierarchy of research designs." New England Journal of Medicine **342**: 1887-1892.

Conway, S. P. (1997). "Recombinant human DNase (rhDNase) in cystic fibrosis: is it cost effective?" Archives of Disease in Childhood **77**: 1-3.

Cook, T. and D. Campbell (1979). Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago, Rand McNally College Publishing Company.

Cooper, L. (2009) " Charities call for "public-researcher partnership" in sharing electronic patient data". British Medical Journal. **338**: b856.

Corey, M., L. Edwards, H. Levison, et al. (1997). "Longitudinal analysis of pulmonary function decline in patients with cystic fibrosis." The Journal of Pediatrics **131**: 809-814.

Cotes, J. E. (1993). Lung function: assessment and application in medicine. Oxford, Balckwell Scientific Publications.

CRD (2001). Undertaking Systematic Reviews of Research on Effectiveness. York, Centre for Reviews and Dissemination (CRD), York.

Criswell, L., K. Saag, K. Sems, et al. (1998). "Moderate-term, low-dose corticosteroids for rheumatoid arthritis (Cochrane Review)". In: The Cochrane Library. Oxford, Update Software.

Cucherat, M., E. Bonnefoy and G. Tremeau (1999). "Primary angioplasty versus intravenous thrombolysis for acute myocardial infarction (Cochrane Review)". In: The Cochrane Library. Oxford, Update Software.

Currie, C. and T. MacDonald (2000). "Use of routine healthcare data in safe and cost-effective drug use." Drug Safety **22**: 97-102.

Cutting, G. R. (2000). "Phenotype-genotype relationships". In: M. Hodson and D. Geddes: Cystic Fibrosis. London, Arnold: 49-60.

Dapena, J., C. Ramos, M. Martinez, et al. (1997). Aerosolised rhDNase improves pulmonary function in CF patients, but cannot stop lung impairment: two years of experience with rhDNase treatment. European Working Group for Cystic Fibrosis Meeting, Davos, Switzerland.

Davies, J., M. T. Trindade, C. Wallis, et al. (1997). "Retrospective review of the effects of rhDNase in children with cystic fibrosis." Pediatric Pulmonology **23**: 243-248.

De Vuyst, J., D. Baran, J. Bertrand, et al. (2000). CF patients treated during 18 months with rhDNase: Belgian multi-centre study. XIIIth International Cystic Fibrosis Congress, Stockholm, Sweden.

Deeks J., J. Dinnes, R. D'Amico, et al. (2003). "Evaluating Non-Randomised Intervention Studies." Health Technology Assessment **7** (27).

Delaisi, B., C. Grosskopf, E. Reignault, et al. (1998). "Epidemiologic international registry of cystic fibrosis: Comparison of French data with European data for 1995. [French]." Archives de Pediatrie **5**: 384-388.

Deneuville, E., J. Dassonville, A. Dabadie, et al. (1997). Two years follow-up of CF patients treated with rhDNase in the CF centre of Rennes (France). European Working Group for Cystic Fibrosis Meeting, Davos, Switzerland.

Department of Health (2001a). Building the Information Core: Protecting and Using Confidential Patient Information - A Strategy for the NHS. London, Department of Health, Information Policy Unit.

Department of Health (2001b). Research Governance Framework for Health and Social Care. London, Department of Health.

Department of Health (2001c). Research Governance Framework for Health and Social Care. London, Department of Health.

Department of Health (2003). Confidentiality: NHS Code of Practice. London, Department of Health.

Department of Health (2009). The Handbook to the NHS Constitution for England. London, Department of Health.

Dodd, M., C. Haworth, A. Webb, et al. (1998). "The medical and economic consequences of non-compliance in clinical trials and in primary care." Thorax **53**: A18.

Dodge, J. A., S. Morison, P. A. Lewis, et al. (1997). "Incidence, population, and survival of cystic fibrosis in the UK, 1968-95. UK Cystic Fibrosis Survey Management Committee." Archives of Disease in Childhood **77**: 493-496.

Downs, S. and N. Black (1997). "The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions." Journal of Epidemiology and Community Health **52**: 377-384.

Drummond, M. F., B. J. O'Brien, G. L. Stoddart, et al. (1997). Methods for the economic evaluation of health care programmes. Oxford, Oxford University Press.

Eddy, D., V. Hasselblad and R. Shachter (1992). Meta-Analysis by the Confidence Profile Method. London, Academic Press.

Eggleston, A., A. Wigerinck and A. Haycox (1996). "Outcome research in GORD: retrospective analysis of prescription data (Mediplus UK) on cisapride, ranitidine and omeprazole." Gut **39**(Suppl 3): A106.

Eggleston, A., A. Wigerinck, S. Huijghebeart, et al. (1998). "Cost effectiveness of treatment for gastro-oesophageal reflux disease in clinical practice: a clinical database analysis." Gut **42**: 13-16.

Eisenberg, J. D., M. L. Aitken, H. L. Dorkin, et al. (1997). "Safety of repeated intermittent courses of aerosolized recombinant human deoxyribonuclease in patients with cystic fibrosis." Journal of Pediatrics **131**: 118-124.

Elborn, J. S., D. J. Shale and J. R. Britton (1991). "Cystic fibrosis: current survival and population estimates to the year 2000." Thorax **46**: 881-885.

Elm, E., D. Altman, M. Egger, et al. (2007). "Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies." The Lancet **370**: 1453-1457.

Elwood, M. (1998). Critical Appraisal of Epidemiological Studies and Clinical Trials. Oxford, Oxford University Press.

Empire Rheumatism Council (1955). "Multi-centre controlled trial comparing cortisone acetate and acetyl salicylic acid in the long-term treatment of rheumatoid arthritis." Annals of Rheumatic Diseases **14**: 353-370.

Eng, P. A., J. Morton, J. A. Douglass, et al. (1996). "Short-term efficacy of ultrasonically nebulized hypertonic saline in cystic fibrosis." Pediatric Pulmonology **21**: 77-83.

EPOC (1998). The Data Collection Checklist. Aberdeen, Cochrane Effective Practice and Organisation of Care Review Group (EPOC).

European Parliament and Council (1995). "Directive 95/46/EC of the of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data." official Journal of the European Communities **23 November 1995**(L.281): 31-50.

Every, N., P. Frederick, M. Robinson, et al. (1999). "A comparison of the National Registry of Myocardial Infarction 2 with the Cooperative Cardiovascular Project." Journal of the American College of Cardiology **33**: 1886-1894.

Fedson, D. S., A. Wajda, J. P. Nicol, et al. (1993). "Clinical effectiveness of influenza vaccination in Manitoba." Journal of the American Medical Association **270**: 1956-1961.

Feit, F., M. Mori Brooks, G. Sopko, et al. (1999). "Long-term clinical outcome in the bypass angioplasty revascularization investigation registry - comparison with the randomized trial." Circulation **101**: 2795-2802.

Feldman, B., E. Wang, A. Willan, et al. (2001). "The randomized placebo-phase design for clinical trials." Journal of Clinical Epidemiology **54**: 550-557.

Fistein, J. (2008) Patient Consent Approach. NHS Connecting for Health - Research Capability Programme. London, Department of Health (PD15) NPFIT-RCP-IG-0001.

FitzSimmons, S. C. (1994). "The changing epidemiology of cystic fibrosis." Current Problems in Pediatrics **24**: 171-179.

Frangolias, D. D., E. Mahenthiralingam, S. Rae, et al. (1999). "Burkholderia cepacia in Cystic Fibrosis . Variable Disease Course." American Journal of Respiratory & Critical Care Medicine **160**: 1572-1577.

Fuchs, H. J., D. S. Borowitz, D. H. Christiansen, et al. (1994). "Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. The Pulmozyme Study Group." New England Journal of Medicine **331**: 637-642.

Furuya, M. E. Y., J. L. Lezana-Fernandez, M. H. Vargas, et al. (2001). "Efficacy of human recombinant DNase in pediatric patients with cystic fibrosis." Archives of Medical Research **32**: 30-34.

Gable, C. B., S. S. Holzer, L. Engelhart, et al. (1990). "Pneumococcal vaccine. Efficacy and associated cost savings." Journal of the American Medical Association **264**: 2910-2915.

Galmiche, J., P. Barthelemy and B. Hamelin (1997). "Treating the symptoms of gastro-oesophageal reflux disease: a double-blind comparison of omeprazole and cisapride." Alimentary Pharmacological Therapy **11**: 765-773.

Gan, K. H., W. P. Geus, W. Bakker, et al. (1995). "Genetic and clinical features of patients with cystic fibrosis diagnosed after the age of 16 years." Thorax **50**: 1301-1304.

GAO (1992). Cross Design Synthesis - A New Strategy for Medical Effectiveness Research. Washington, United States General Accounting office.

Garcia, E., J. Elizaga, N. Perez-Castellano, et al. (1999). "Primary angioplasty versus thrombolysis in the anterior myocardial infarction." Journal of the American College of Cardiology **33**: 605-611.

Garcia Rodriguez, L. A. and A. Ruigomez (1999). "Secondary prevention of upper gastrointestinal bleeding associated with maintenance acid-suppressing treatment in patients with peptic ulcer bleed." Epidemiology **10**: 228-232.

Geddes, D. M. and P. L. Shah (1999). "Where we are now with rhDNase." Lancet **353**: 1727.

Ghani, A. C., W. E. Henley, C. A. Donnelly, et al. (2001). "Comparison of the effectiveness of non-nucleoside reverse transcriptase inhibitor-containing and protease inhibitor-containing regimens using observational databases." AIDS **15**: 1133-1142.

Giralt, S., R. Szydlo, J. M. Goldman, et al. (2000). "Effect of short-term interferon therapy on the outcome of subsequent HLA-identical sibling bone marrow transplantation for chronic myelogenous leukemia: an analysis from the international bone marrow transplant registry." Blood **95**: 410-415.

Gliklich, R.E., N.A. Dreyer (eds.) (2007). Registries for Evaluating Patient Outcomes: A User's Guide. AHRQ Publication No. 07-EHC001-1. Rockville, MD: Agency for Healthcare Research and Quality. April 2007.

GMC (1998). Seeking Patients' Consent: the Ethical Considerations. London, General Medical Council.

GMC (2000). Confidentiality: Protecting and Providing Information. London, General Medical Council.

GMC (2002). Research: The Role and Responsibilities of Doctors. London, General Medical Council.

Goldstein, R. E., M. Andrews, W. J. Hall, et al. (1996). "Marked reduction in long-term cardiac deaths with aspirin after a coronary event. Multicenter Myocardial Ischemia Research Group." J Am Coll Cardiol **28**: 326-330.

Goncz, K. K., K. Kunzelmann, Z. Xu, et al. (1998). "Targeted replacement of normal and mutant CFTR sequences in human airway epithelial cells using DNA fragments." Human Molecular Genetics **7**: 1913-1919.

Graham, N. M., S. L. Zeger, L. P. Park, et al. (1991). "Effect of zidovudine and Pneumocystis carinii pneumonia prophylaxis on progression of HIV-1 infection to AIDS. The Multicenter AIDS Cohort Study." The Lancet **338**: 265-269.

Gray-Donald, K. and M. Kramer (1988). "Causality inference in observational vs. experimental studies." American Journal of Epidemiology **127**: 885-892.

Green, S. and D. Byar (1984). "Using observational data from registries to compare treatments: the fallacy of omnimetrics." Statistics in Medicine **3**: 361-370.

Greenhouse, J. and K. Kelleher (2005). "Thinking outside the (black) box: Anti-depressants, suicidality and research synthesis." Pediatrics **116**: 231-233.

Greenland, S. (1987). "Quantitative methods in the review of epidemiologic literature." Epidemiologic Reviews **9**: 1-30.

Grines, C., K. Browne, J. Marco, et al. (1993). "A comparison of immediate angioplasty with thrombolytic therapy for acute myocardial infarction." New England Journal of Medicine **328**: 673-679.

GUSTO IIb (1997). "A clinical trial comparing primary coronary angioplasty with tissue plasminogen activator for acute myocardial infarction (The Global Use of Strategies to Open Occluded Coronary Arteries in Acute Coronary Syndromes angioplasty substudy investigators)." New England Journal of Medicine **336**: 1621-1628.

Hallerbäck, B., H. Glise, B. Johansson, et al. (1998). "Gastro-oesophageal reflux symptoms - clinical findings and effect of ranitidine treatment." Euruopean Journal of Surgery **583**: S6-13.

Harms, H. K., E. Matouk, G. Tournier, et al. (1998a). "Multicenter, open-label study of recombinant human DNase in cystic fibrosis patients with moderate lung disease. DNase International Study Group." Pediatric Pulmonology **26**: 155-161.

Harms, H. K., J. Navarro, M. Hodson, et al. (1998b). The Epidemiologic Registry of Cystic Fibrosis (ERCF). What differences exist between countries and centres? European CF Society Meeting, Berlin, Germany.

Harris, E., R. Emkey, J. Nichols, et al. (1983). "Low dose prednisone therapy in rheumatoid arthritis: A double blind study." Journal of Rheumatology **10**: 713-721.

Hartwell, D., J. Colquitt, E. Loveman, et al. (2005) "Clinical effectiveness and cost-effectiveness of immediate angioplasty for acute myocardial infarction: systematic review and economic evaluation" Health Technology Assessment **17**:1-114.

Hartz, A., S. Bentler, M. Charlton, et al. (2005). "Assessing observational studies of medical treatments." Emerging Themes in Epidemiology **2**: 8. (http://www.ete-online.com/content/2/1/8)

Hatlebakk, J., A. Hyggen, P. Madsen, et al. (1999). "Heartburn treatment in primary care: randomised, double-blind study for 8 weeks." British Journal of Medicine **319**: 550-553.

Hawley, D. and F. Wolfe (1991). "Are the results of controlled clinical trials and observational studies of second line therapy in rheumatoid arthritis valid and generalizable as measures of rheumatoid arthritis outcome: analysis of 122 studies." Journal of Rheumatology **18**: 1008-1014.

Haynes, C., G. Cook and M. Jones (2007) "Legal and ethical considerations in processing patient-identifiable data without patient consent: lessons learnt from developing a disease registe." Journal of Medical Ethics **33**: 302-307.

Hennekens, C. and J. Buring (1987). Epidemiology in Medicine. Boston/Toronto, Little, Brown and Company.

Henry, R. L., P. G. Gibson, K. Carty, et al. (1998). "Airway inflammation after treatment with aerosolized deoxyribonuclease in cystic fibrosis." Pediatric Pulmonology **26**: 97-100.

Heuckmann, L., P. Kuster and J. Uekotter (1999). "Long-term effect of DNase in cystic fibrosis. [German]." Monatsschrift für Kinderheilkunde **147**: 573-576.

Heudebert, G. R., J. Van Ruiswyk, J. Hiatt, et al. (1993). "Combination drug therapy for hypercholesterolemia. The trade-off between cost and simplicity." Archives of Internal Medicine **153**: 1828-1837.

Hill, A.B. (1965). "The environment and disease: association or causation?" Proceedings of the Royal Society of Medicine **58**: 295-300.

Hlatky, M., R. Califf, F. Harrell, et al. (1988). "Comparison of predictions based on observational data with the result of randomized controlled trials of coronary artery bypass surgery." Journal of the American College of Cardiology **11**: 237-245.

Hodson, M., M. Rainisio, S. McKenzie, et al. (1998a). The Epidemiologic Registry of Cystic Fibrosis (ERCF). In 1998 - Where are we? European CF Society Meeting, Berlin, Germany.

Hodson, M., M. Rainisio, S. McKenzie, et al. (1998b). The European Epidemiologic Registry of Cystic Fibrosis (ERCF). Does pulmozyme make any difference? European CF Society Meeting, Berlin, Germany.

Hodson, M., M. Rainisio, S. McKenzie, et al. (1999). "The European Epidemiological Registry of Cystic Fibrosis (ERCF). Does pulmozyme make any difference?" The Netherlands Journal of Medicine **54**: S22-23.

Hodson, M., S. McKenzie, H. Harms, C. Koch, G. Mastella, J. Navarro, B. Strandvik (2003). "Dornase alfa in the treatment of cystic fibrosis in Europe: a report from the Epidemiologic Registry of Cystic Fibrosis." Pediatric Pulmonology **36**: 427-432.

Hogel, J., H. Walach and W. Gaus (1994). "Change-to-open-label design: Proposal and discussion of a new design for clinical parallel-group double-masked trials." Arzneimittelforschung **14**: 97-99.

Holloway, J. and D. Schocken (1988). "CASS in retrospect: lessons from the randomized cohort and registry." The American Journal of Medical Sciences **295**: 424-432.

Homik, J., A. Cranney, B. Shea, et al. (1998). "Bisphosphonates for steroid induced osteoporosis (Cochranre Review)". In: The Cochrane Library. Oxford, Update Software.

Huang, X., A. G. Hartzema, R. H. Raasch, et al. (1999). "Economic assessment of three antimicrobial therapies for uncomplicated urinary tract infection in women." Clinical Therapeutics **21**: 1578-1588.

Hubbard, R. C., N. G. McElvaney, P. Birrer, et al. (1992). "A preliminary study of aerosolized recombinant human deoxyribonuclease I in the treatment of cystic fibrosis." New England Journal of Medicine **326**: 812-815.

Hughes, D., A. Bagust, A. Haycox, et al. (2001). "Accounting for Noncompliance in Pharmacoeconomic Evaluations." Pharmacoeconomics **19**: 1185-1197.

Huston, P. and C. D. Naylor (1996). "Health service research: reporting on studies using secondary data sources." Canadian Medical Association Journal **155**: 1697-1702.

IBMTR (1989). "Effect of methotrexate on relapse after bone-marrow transplantation for acute lymphoblastic leukaemia." The Lancet(March 11): 535-537.

Information Commissioner (2001). Data Protection Act 1998 - Legal Guidance. Wilmslow, Information Commissioner.

Information Commissioner (2002). Use and Disclosure of Health Data - Guidance on the Application of the Data Protection Act 1998. Wilmslow, Information Commissioner.

Innes, J. A. (1998). "DNase in cystic fibrosis: the challenge of assessing response and maximising benefit." Thorax **53**: 1003-1004.

Ioannidis, J., A.-B. Haidich, M. Pappa, et al. (2001). "Comparison of evidence of treatment effects in randomized and nonrandomized studies." Journal of the American Medical Association **286**: 821-830.

Iskedjian, M. and T. Einarson (1998). "Meta-analyses of cisapride, omeprazole and ranitidine in the treatment of GORD: implications for treating patient subgroups." Clinical Drug Investigation **16**: 9-18.

Jadad, A., A. Moore, D. Carrol, et al. (1996). "Assessing the quality of reports of randomized clinical trials: is blinding necessary?" Controlled Clinical Trials **17**: 1-12.

Johnson, C. A., S. M. Butler, M. W. Konstan, et al. (1999a). "Estimating effectiveness in an observational study: a case study of dornase alfa in cystic fibrosis. The Investigators

and Coordinators of the Epidemiologic Study of Cystic Fibrosis." The Journal of Pediatrics **134**: 734-739.

Johnson, J. A., M. A. Connolly, P. Jacobs, et al. (1999b). "Cost of care for individuals with cystic fibrosis: a regression approach to determining the impact of recombinant human DNase." Pharmacotherapy **19**: 1159-1166.

Jones, J. (2000). "Doctors suspended in child health inquiry." British Medical Journal **320**: 9.

Jüni, P., D. Altman and M. Egger (2001). "Assessing the quality of randomised controlled trials". In: M. Egger, G. Davey Smith and D. Altman: Systematic Reviews in Health Care: Meta-Analysis in Context. London, BMJ Publishing Group: 87-108.

Kaptchuk, T. (2001). "the double-blind, randomized, placebo-controlled trial: Gold standard or golden calf?" Journal of Clinical Epidemiology **54**: 541-549.

Kearney, C. E. and C. E. Wallis (2003). "Deoxyribonuclease for cystic fibrosis (Cochrane Review)". In: The Cochrane Library, Issue 2. Oxford, Update Software.

Kerem, B., J. J. Reisman, M. Corey, et al. (1992). "Prediction of mortality in patients with cystic fibrosis." New England Journal of Medicine **326**: 1187-1191.

Kherani, R., A. Papaioannou and J. Adachi (2002). "Long-term tolerability of the bisphosphonates in postmenopausal osteoporosis: a comparative review." Drug Safety **25**: 781-790.

Kirwan, J. (1995). "The effect of glucocorticoids on joint destruction in rheumatoid arthritis (Arthritis and Rheumatism Council Low Dose Glucocorticoid Study Group)." **333**: 142-146.

Klawansky, S., A. Antczak-Bouckoms, J. Barr, et al. (1991a). "Using medical registries and data sets for technology assessment - an overview of seven case studies." International Journal of Technology Assessment in Health Care **7**: 194-199.

Klawansky, S., E. Burdick, M. Adams, et al. (1991b). "Use of the SEER cancer registry for technology assessment." International Journal of Technology Assessment in Health Care **7**: 134-142.

Klungel, O., E. Martens, B. Psaty, et al. (2004). "Methods to assess intended effects of drug treatment in observational studies are reviewed." Journal of Clinical Epidemiology **57**: 1223-1231.

Knudson, R., M. Lebowitz, C. Holberg, et al. (1983). "Changes in the normal expiry flow-volume curve with groth and aging." American Review of Respiratory Disease **127**: 725-734.

Koch, C., H. Cuppens, M. Rainisio, et al. (2001). "European Epidemiologic Registry of Cystic Fibrosis (ERCF): comparison of major disease manifestations between patients with different classes of mutations." Pediatric Pulmonology **31**: 1-12.

Koch, C., S. G. McKenzie, H. Kaplowitz, et al. (1997). "International practice patterns by age and severity of lung disease in cystic fibrosis: data from the Epidemiologic Registry of Cystic Fibrosis (ERCF)." Pediatric Pulmonology **24**: 147-154; discussion 159-161.

Koch, C., M. Rainisio, H. Harms, et al. (1999). "Risk factors of importance for long-term survival of patients with cystic fibrosis." The Netherlands Journal of Medicine **54**(22).

Koch, C., M. Rainisio, U. Madessani, et al. (2000). Impact of diabetes mellitus on lung function. XIIIth International Cystic Fibrosis Congress, Stockholm, Sweden.

Konstan, M., P. J. Byard, C. L. Hoppel, et al. (1995). "Effect of high-dose ibuprofen in patients with cystic fibrosis." New England Journal of Medicine **332**: 848-854.

Korn, E. and S. Baumrind (1991). "Randomised clinical trials with clinician-preferred treatment." The Lancet **337**: 149-152.

Krumholz, H. M., Y. T. Chen, Y. Wang, et al. (2001). "Aspirin and angiotensin-converting enzyme inhibitors among elderly survivors of hospitalization for an acute myocardial infarction." Archives of Internal Medicine **161**: 538-544.

Krumholz, H. M., J. Hennen, P. M. Ridker, et al. (1998). "Use and effectiveness of intravenous heparin therapy for treatment of acute myocardial infarction in the elderly." Journal of the American College of Cardiology **31**: 973-979.

Krumholz, H. M., M. J. Radford, E. F. Ellerbeck, et al. (1995). "Aspirin in the treatment of acute myocardial infarction in elderly Medicare beneficiaries. Patterns of use and outcomes." Circulation **92**: 2841-2847.

Kuhn, L., E. J. Abrams, J. Weedon, et al. (2000). "Disease progression and early viral dynamics in human immunodeficiency virus-infected children exposed to zidovudine during prenatal and perinatal periods." Journal of Infectious Diseases **182**: 104-111.

Kunz, R. and A. Oxman (1998). "The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials." British Medical Journal **317**: 1185-1190.

Labin, S. (2007) "Research synthesis - towards broad-based evidence." In: N. Smith and P. Brandon: Fundamental Issues in Evaluation Guildford Press: 89-110.

Lai, H. C., M. R. Kosorok, S. A. Sondel, et al. (1998). "Growth status in children with cystic fibrosis based on the National Cystic Fibrosis Patient Registry data: Evaluation of various criteria used to identify malnutrition." Journal of Pediatrics **132**: 478-485.

Lang, S., A. Hansen, B. Thorsteinsson, et al. (1995). "Glucose tolerance in patients with CF: five year prospective study." British Medical Journal **311**: 655-659.

Lang, S., B. Thorsteinsson, C. Lund-Anderson, et al. (1994). "Diabetes mellitus in Danish cystic fibrosis patients: prevalence and late diabetic complications." Acta Paediatrica **83**: 72-77.

Lasagna, L. (1974). "A plea for the "naturalistic" study of medicines." European Journal of Clinical Pharmacology **7**: 153-154.

Laube, B. L., R. M. Auci, D. E. Shields, et al. (1996). "Effect of rhDNase on airflow obstruction and mucociliary clearance in cystic fibrosis." American Journal of Respiratory & Critical Care Medicine **153**: 752-760.

Lawrenson, R. A. and J. W. Logie (2001). "Antibiotic failure in the treatment of urinary tract infections in young women." J Antimicrob Chemother **48**: 895-901.

Ledergerber, B., J. von Overbeck, M. Egger, et al. (1994). "The Swiss HIV Cohort Study: rationale, organization and selected baseline characteristics." Sozial- und Präventivmedizin **39**: 387-394.

Ledson, M. J., Z. Wahbi, R. P. Convery, et al. (1998). "Targeting of dornase alpha therapy in adult cystic fibrosis." Journal of the Royal Society of Medicine **91**: 360-364.

Lewis, N., J. Patwell and B. Briesacher (1993). "The role of insurance claims databases in drug therapy outcomes research." Pharmacoeconomics **4**: 323-330.

Lewis, P. (2000). "The epidemiology of cystic fibrosis". In: M. Hodson and D. Geddes: Cystic Fibrosis. London, Arnold: 13-26.

Lewis, P. (2001). How good are CF registries? ECFC Symposium, Vienna, ECFSOC.

Lewis, P. A., S. Morison, J. A. Dodge, et al. (1999). "Survival estimates for adults with cystic fibrosis born in the United Kingdom between 1947 and 1967. The UK Cystic Fibrosis Survey Management Committee." Thorax **54**: 420-422.

Lewsey, J., A. Leyland, G. Murray, et al. (2000). "Using Routine Data to Complement and enhance the Results of Randomised Controlled Trials." Health Technology Assessment **4**(22).

Littlewood, J. M. and S. Wolfe (2000). "Growth, development and nutrition". In: M. Hodson and D. Geddes: Cystic Fibrosis. London, Arnold: 243-260.

Lowrance, W. (2006). Access to Collections of Data and Materials for Health Research. London, Medical Research Council and WellcomeTrust.

Lucidi, V., P. Rosati, M. Realacci, et al. (1997). A longitudinal study of pulmonary function in a CF population. European Working Group for Cystic Fibrosis Meeting, Davos, Switzerland.

Lundgren, J. D., A. N. Phillips, C. Pedersen, et al. (1994). "Comparison of long-term prognosis of patients with AIDS treated and not treated with zidovudine. AIDS in Europe Study Group." Journal of the American Medical Association **271**: 1088-1092.

MacLehose, R., B. Reeves, I. Harvey, et al. (2000). "A Systematic Review of Comparisons of Effect Sizes Derived from Randomised and Non-Randomised Studies." Health Technology Assessment **4**(34).

MacMahon, S. and R. Collins (2001). "Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies." The Lancet **357**: 455-462.

Magee, L., S. Bull, G. Koren, et al. (2001). "The generalizability of trial data: a comparison of beta-blocker trial participants with a prospective cohort of women taking beta-blockers in pregnancy." European Journal of Obstetrics and Gynecology **94**: 205-210.

Mastella, G., M. Hodson, C. Koch, et al. (1999). "Factors affecting velocity of lung function deterioration in ERCF cystic fibrosis over 2-4 years follow-up." The Netherlands Journal of Medicine **54**.

Mastella, G., U. Madessani, M. Rainisio, et al. (2000). Epidemiological survey of bacteria isolated from respiratory tract in European CF patients. XIIIth International Cystic Fibrosis Congress, Stockholm, Sweden.

McCoy, K., S. Hamilton and C. Johnson (1996). "Effects of 12-week administration of dornase alfa in patients with advanced cystic fibrosis lung disease. Pulmozyme Study Group." Chest **110**: 889-895.

McDougall, R., J. Sibley, M. Haga, et al. (1994). "Outcome in patients with rheumatoid arthritis receiving prednisone compared to matched controls." Journal of Rheumatology **21**: 1207-1213.

McIntyre, A. M. (1999). "Dornase alpha and survival of patients with cystic fibrosis." Hospital Medicine **60**: 736-739.

Menzin, J., G. Oster, L. Davies, et al. (1996). "A multinational economic evaluation of rhDNase in the treatment of cystic fibrosis." International Journal of Technology Assessment in Health Care **12**: 52-61.

Metcalfe, C., R. Martin, S. Noble, et al. (2008) "Low risk research using routinely collected identifiable health information without informed consent: encounters with the Patient Information Advisory Group" Journal of Medical Ethics **34**: 37-40.

Michels, K.B. and Yusuf S. (1995) "Does PTCA in acute myocardial infarction affect mortality and reinfarction rates? A quantitative overview (meta-analysis) of the randomised clinical trials." Circulation **91**: 476-485.

Milla, C. E. (1998). "Long term effects of aerosolised rhDNase on pulmonary disease progression in patients with cystic fibrosis." Thorax **53**: 1014-1017.

Million, R., P. Poole, J. Kellgren, et al. (1984). "Long-term study management of rheumatoid arthritis." The Lancet(April 14): 812-816.

Moore, R. D., J. Hidalgo, B. W. Sugland, et al. (1991). "Zidovudine and the natural history of the acquired immunodeficiency syndrome." New England Journal of Medicine **324**: 1412-1416.

Morgan, W. J., S. M. Butler, C. A. Johnson, et al. (1999). "Epidemiologic study of cystic fibrosis: Design and implementation of a prospective, multicenter, observational study of patients with cystic fibrosis in the U.S. and Canada." Pediatric Pulmonology **28**: 231-241.

Morison, S., J. A. Dodge, T. J. Cole, et al. (1997). "Height and weight in cystic fibrosis: a cross sectional study. UK Cystic Fibrosis Survey Management Committee." Archives of Disease in Childhood **77**: 497-500.

Morris, H. F., S. Ochi and S. Winkler (2000). "Implant survival in patients with type 2 diabetes: placement to 36 months." Ann Periodontol **5**(1): 157-165.

Mosis, G. (2006) Randomized Database Studies: A Combination of the Strenths of Observational Studies and Randomized Clinical Trials? PhD thesis, Erasmus University, Rotterdam. http://publishing.eur.nl/ir/repub/asset/8054/RDS_G_Mosis.pdf (accessed on 18 October 2009)

Motheral, B. and K. Fairman (1997). "The use of claims databases for outcomes research: rationale, challenges, and strategies." Clinical Therapeutics **19**: 346-366.

Mowatt, G., D. Bower, J. Brebner, et al. (1997). "When and how to assess fast-changing technologies: a comparative study of medical applications of four generic technologies." Health Technology Assessment **1**(14).

MRC (1999) "Personal Information in Medical Research: Draft MRC

Guidelines." London, Medical Research Council, Working Group on Personal Information in Medical Research. www.mrc.ac.uk/per info.html (accessed on 29 August 2000).
MRC (2000). Personal Information in Medical Research. London, Medical Research Council.
Nasr, S. Z., L. R. Kuhns, R. W. Brown, et al. (2001). "Use of computerized tomography and chest x-rays in evaluating efficacy of aerosolized recombinant human DNase in cystic fibrosis patients younger than age 5 years: a preliminary study." Pediatric Pulmonology **31**: 377-382.
National Coordinating Centre for Health Technology Assessment (2009) "About the HTA programme." http://www.ncchta.org/ (accessed on 26 March 2009).
National Institute for Clinical Excellence (2004) "Dyspepsia - Management of Dyspepsia in Adults in Primary Care" Clinical Guideline 17. London, National Institute for Clinical Excellence.
National Institute for Health Research (2009) "About the Health Technology Assessment Programme" http://www.hta.nhs.uk/publicationspdfs/Infoleaflets/GeneralLeaflet.pdf (accessed on 3 October 2009)
National Library of Medicine (2001) "Abridged Index Medicus (AIM) Journal Titles." http://www.nlm.nih.gov/bsd/aim.html (accessed on 14 July 2003).
Navarro, J., O. Mouterde, H. Harms, et al. (1999). "Influence of screening on medical status of CF patients. Data from the ERCF." The Netherlands Journal of Medicine **54**.
Navarro, J., M. Rainisio, H. K. Harms, et al. (2001). "Factors associated with poor pulmonary function: cross-sectional analysis of data from the ERCF. European Epidemiologic Registry of Cystic Fibrosis." European Respiratory Journal **18**: 298-305.
Navarro, J., M. Rainisio, S. McKenzie, et al. (2000). Evolution of weight for age Z score and medical conditions in ERCF patients. XIIIth International Cystic Fibrosis Congress, Stockholm, Sweden.
Newgard, C., J. Hedges, M. Arthur, R. Mullins (2004). "Advanced statistics: The Propensity Score - A method for estmating treatment effect in observational research" Academic Emergency Medicine **11**: 953-961.
Newman, J. J., C. D. Bayliff, B. D. Lyttle, et al. (2001). "The role of dornase alfa in the treatment of severe cystic fibrosis." Canadian Journal of Public Health **54**: 40-42.
NHS Centre for Reviews and Dissamination (2001). Understanding Systematic Reviews of Research on Effectiveness. York.
NHS Executive (1999a). Caldicott Guardians. Leeds, NHS Executive.
NHS Executive (1999b). For the Record. Leeds, NHS Executive.
NHS Executive (1999c). Protecting and Using Patient Information--A Manual for Caldicott Guardians. Leeds, NHS Executive.
NHS Executive (2000). Data Protection Act 1998: Protection and Use of Patient Information. Leeds, NHS Executive.
NHS Information Authority (2002). Caring for Information - Model for the Future. London, NHS Information Authority.
Nichol, K. L., L. Baken and A. Nelson (1999a). "Relation between influenza vaccination and outpatient visits, hospitalization, and mortality in elderly persons with chronic lung disease." Annals of Internal Medicine **130**: 397-403.
Nichol, K. L., L. Baken, J. Wuorenma, et al. (1999b). "The health and economic benefits associated with pneumococcal vaccination of elderly persons with chronic lung disease." Archives of Internal Medicine **159**: 2437-2442.
Nichol, K. L., K. L. Margolis, J. Wuorenma, et al. (1994). "The efficacy and cost effectiveness of vaccination against influenza among elderly persons living in the community." New England Journal of Medicine **331**: 778-784.
Nordin, J., J. Mullooly, S. Poblete, et al. (2001). "Influenza vaccine effectiveness in preventing hospitalizations and deaths in persons 65 years or older in Minnesota, New York, and Oregon: data from 3 health plans." Journal of Infectious Diseases **184**: 665-670.
Nousia-Arvanitakis, S., M. Xefteri, A. Galli-Tsinopoulou, et al. (1997). Effect of 12-month treatment with recombinant human DNase on respiratory function in cystic fibrosis. European Working Group for Cystic Fibrosis Meeting, Davos, Switzerland.

Nowakowska, A., D. Sands, T. Orlik, et al. (1997). Long-term rhDNase therapy in children with CF. European Working Group for Cystic Fibrosis Meeting, Davos, Switzerland.

Nuijten, M. (1998). "The selection of data sources for use in modelling studies." Pharmacoeconomics **13**: 305-316.

Ollendorf, D. A., L. J. McGarry, M. L. Watrous, et al. (2000). "Use of rhDNase therapy and costs of respiratory-related care in patients with cystic fibrosis." Annals of Pharmacotherapy **34**: 304-308.

Olschewski, M. and H. Scheurlen (1985). "Comprehensive cohort study: an alternative to randomized consent design in a breat preservation trial." Methods Inf Med **24**: 131-134.

Orozco, L., L. Gonzalez, M. Chavez, et al. (2001). "XV-2c/KM-19 haplotype analysis of cystic fibrosis mutations in Mexican patients." American Journal of Medical Genetics **102**: 277-281.

Oster, G., D. M. Huse, M. J. Lacey, et al. (1995). "Effects of recombinant human DNase therapy on healthcare use and costs in patients with cystic fibrosis." Annals of Pharmacotherapy **29**: 459-464.

Ottenbacher, K (1992). " Impact of random assignment on study outcome: an empirical examination." Controlled Clinical Trials **13**: 50-61.

Paredes, R., A. Mocroft, O. Kirk, et al. (2000). "Predictors of virological success and ensuing failure in HIV-positive patients starting highly active antiretroviral therapy in Europe: results from the EuroSIDA study." Arch Intern Med **160**(8): 1123-1132.

Pauker, S. and J. Kassirer (1987). "Decision analysis." New England Journal of Medicine **316**: 250-258.

Penketh, A. R., A. Wise, M. B. Mearns, et al. (1987). "Cystic fibrosis in adolescents and adults." Thorax **42**: 526-532.

Perras, C. and N. Otten (1996). Pulmozyme: clinical and economic impacts, Canadian Coordinating office for Health Technology Assessment (CCOHTA).

Peterson, L. R., N. C. Chandra, W. J. French, et al. (1999). "Reperfusion therapy in patients with acute myocardial infarction and prior coronary artery bypass graft surgery (National Registry of Myocardial Infarction-2)." American Journal of Cardiology **84**: 1287-1291.

Pethica, B. D., A. Penrose, D. MacKenzie, et al. (1998). "Comparison of potency of inhaled beclomethasone and budesonide in New Zealand: retrospective study of computerised general practice records." British Medical Journal **317**: 986-990.

Petitti, D. (1994). Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine. Oxford, Oxford University Press.

Peto, R., R. Collins and R. Gray (1995). "Large-scale randomized evidence: large, simple trials and overviews of trials." Journal of Clinical Epidemiology **48**: 23-40.

Phillips, B., C. Ball, D. Sackett, et al. (2001a) "Levels of evidence and grades of recommendations." Oxford Centre for Evidence-Based Medicine. http://minerva.minervation.com/cebm/docs/levels.html (accessed on 5 November 2002).

Phillips, J., J. Bunting, P. Masel, et al. (2001b). "Pulmozyme adherence in an adult cystic fibrosis unit." Pediatr Pulmonology Supplement **22**: A407.

Pincus, T. (1988). "Rheumatoid arthirits: disappointing long-term outcomes despite successful short-term clinical trials." Journal of Clinical Epidemiology **41**: 1037-1041.

Pincus, T. (1993). "Limitations of randomized clinical trials to recognize possible advantages of combination therapies in rheumatic diseases." Seminars in Arthritis and Rheumatism **23**(Suppl 1): 2-10.

Pincus, T., S. Marcum and L. Callahan (1992). "Long-term drug therapy for rheumatoid arthritis in seven rheumatology private practices II: second-line drugs and prednisone." Journal of Rheumatology **19**: 1885-1894.

Pitt, P., F. Li, B. Bloom, et al. (1998). "A double-blind placebo controlled study to determine the effects of intermittent cyclical etidronate on bone mineral density in patients on long term corticosteroid treatment." Thorax **53**: 351-356.

Polit, D. and B. Hungler (1991). Nursing Research: Principles and Methods. Philadelphia, J.B. Lippincott Company.

Price, D. B., K. Cargill, S. Wolfe, et al. (1998). "Salmeterol xinafoate: an analysis of outcomes and cost-effectiveness using a primary care database." Respiratory Medicine **92**: 1302-1304.

Quan, J. M., H. A. Tiddens, J. P. Sy, et al. (2001). "A two-year randomized, placebo-controlled trial of dornase alfa in young patients with cystic fibrosis with mild lung function abnormalities." The Journal of Pediatrics **139**: 813-820.

R&D office (2002). Research Governance Framework for Health and Social Care. Belfast, Department of Health, Social Services, and Public Safety.

Rabeneck, L., T. Menke, M. Simberkoff, et al. (2001). "Using the national registry of HIV-infected veterans in research: lessons for the development of disease registries." Journal of Clinical Epidemiology **54**: 1195-1203.

Rabinowitz, J., P. Lichtenberg, Z. Kaplan, et al. (2001). "Rehospitalization rates of chronically ill schizophrenic patients discharged on a regimen of risperidone, olanzapine, or conventional antipsychotics." Am J Psychiatry **158**: 266-269.

Raftery, J., P. Roderick, A. Stevens (2005). "Potential Use of Routine Databases in Health Technology Assessment." Health Technology Assessment **9**(20).

Rahme, E., L. Pilote and J. LeLorier (2002). "Association between naproxen use and protection against acute myocardial infarction." Archives of Internal Medicine **162**: 1111-1115.

Ramsey, B. and T. Boat (1994). "Outcome measures for clinical trials in cystic fibrosis: Summary of a Cystic Fibrosis Foundation Consensus Conference." Journal of Pediatrics **124**: 177-192.

Ramsey, B. W., S. J. Astley, M. L. Aitken, et al. (1993). "Efficacy and safety of short-term administration of aerosolized recombinant human deoxyribonuclease in patients with cystic fibrosis." American Review of Respiratory Disease **148**: 145-151.

Ranasinha, C., B. Assoufi, S. Shak, et al. (1993). "Efficacy and safety of short-term administration of aerosolised recombinant human DNase I in adults with stable stage cystic fibrosis." The Lancet **342**: 199-202.

Ratnalingam, R., N. Vasudev, A. Watson, et al. (2001). Benefit from Pulmozyme use in an adult CF unit 1994-2000: comparison of responders and non-responders. 24th European CF Conference, Vienna, Austria.

Rawlins, M. (2008). "De testimonio: on the evidence for decisions about the use of therapeutic interventions." The Lancet **373**: 2152-2161.

Revicki, D. and L. Frank (1999). "Pharmacoeconomic evaluation in the real world: effectiveness versus efficacy studies." Pharmacoeconomics **15**: 423-434.

Ribichini, F., G. Steffenino, A. Dellavalle, et al. (1998). "Comparison of thrombolytic therapy and primary coronoary angioplasty with liberal stenting for inferior myocardial infarction with precordial ST-segment depression: immediate and long-term results of a randomized study." Journal of the American College of Cardiology **32**: 1687-1694.

Richter, J., D. Peura, S. Benjamin, et al. (2000). "Efficacy of omepazole for the treatment of symptomatic acid reflux disease without esophagitis." Archives of Internal Medicine **160**: 1810-1816.

Rimm, A., J. Barr, M. Horowitz, et al. (1991). "Use of a clinical data registry to evaluate medical technologies - experience from the International Bone Marrow Transplant Registry." International Journal of Technology Assessment in Health Care **7**: 182-193.

Rittenhouse, B. (1995). "Therelevance of searching for effects under a clinical-trial lamppost: a key issue." Medical Decision Making **15**: 348-357.

Rittenhouse, B. and B. J. O'Brien (1996). "Threats to the validity of pharmacoeconomic analyses based on clincal trial data". In: B. Spilker: Quality of Life and Pharmacoeconomics in Clinical Trials. Philadelphia, Lippincott-Raven Publishers: 1215-1223.

Roberts, M., R. D'Agostino, M. Dillon, et al. (1991). "Technology Assessment in the Framingham Heart Study." International Journal of Technology Assessment in Health Care **7**: 156-170.

Robinson, M., A. L. Hemming, C. Moriarty, et al. (2000). "Effect of a short course of rhDNase on cough and mucociliary clearance in patients with cystic fibrosis." Pediatric Pulmonology **30**: 16-24.

Rogers, W., L. Bowlby, N. C. Chandra, et al. (1994). "Treatment of myocardial infarction in the United States (1990 to 1993): observations from the National Registry of Myocardial Infarction." Circulation **90**: 2103-2114.

Romano-Critchley, G. and A. Sommerville (1999). Confidentiality and Disclosure of Health Information. London, British Medical Association.

Rosenbaum, P., D. Rubin (1983) " The central role of the propensity score in observational studies for causal effects." Biometrika **70**:41-55.

Rosenstein, B. J. (2000). "Diagnosis". In: M. Hodson and D. Geddes: Cystic Fibrosis. London, Arnold: 178-188.

Rothman, K. and S. Greenland (1998). Modern Epidemiology. Philadelphia, Lippincott Raven.

Rothman, K. and C.E Wentworth (2003). "Mortality of Cystic Fibrosis Patients Treated with Tobramycin Solution for Inhalation." Epidemiology **14**: 55-59.

Roy, P., G. Vaughan Hudson, B. Vaughan Hudson, et al. (2000). "Long-term survival in Hodkin's disease patients: a comparison of relative survival in patients in trials and those recorded in population-based cancer registries." European Journal of Cancer **36**: 384-389.

Royal College of Physicians Committee on Ethical Issues in Medicine (1999). "Research based on archived information and samples." Journal of the Royal College of Physicians of London **33**: 264-266.

Rubin, D. (1997). "Estimating causal effects from large data sets using propensity scores." Annals of Internal Medicine **127**: 757-763.

Rucker, G. (1989). "A two-stage trial design for testing treatment, self-selection and treatment preference effects." Statistics in Medicine **8**: 477-485.

Sackett, D. (1979). "Bias in analytic research." Journal of Chronic Diseases **32**: 51-63.

Sacristán, J., J. Soto, I. Galende, et al. (1998). "Randomised database studies: a new method t assess drugs' effectiveness?" Journal of Clinical Epidemiology **51**: 713-715.

Sanderson, S., I. Tatt, J. Higgins. (2007) " Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography." International Journal of Epidemiology doi:10.1093/ije/dym018 http://ije.oxfordjournals.org/cgi/reprint/dym018v1

Sands, D., A. Nowakowska, R. Piotrowski, et al. (2000). Long-term dornase alfa therapy in children with CF. XIIIth International Cystic Fibrosis Congress, Stockholm, Sweden.

Santos, C., N. Wizla-Derambure, C. Thumerelle, et al. (2001). Effects of rhDNase on lung function and nutritional status in cystic fibrosis children and adolescents: a four year study. 24th European CF Conference, Vienna, Austria.

Saskatchewan Health (2005). Health Services Databases: Information Document. Regina, SK: Saskatchewan Health (http://www.health.gov.sk.ca/health-databases-info-doc)

Schulenburg, J. v. d., W. Greiner and H. v. d. Hardt (1995). "Socioeconomic evaluation of the effect of rhDNase on the cost of treating infections of the respiratory tract in patients with cystic fibrosis [German]." Medizinische Klinik **90**: 220-224.

Schulman, K., K. Yabroff, J. Kong, et al. (2001). "A claims data approach to defining an episode of care." Pharmacoepidemiology & Drug Safety **10**: 417-427.

Scott, S. F., J. A. Pryor, R. J. Marsden, et al. (2001). Non-responders to rhDNase may respond to therapy with an adaptive aerosol delivery (AAD) system. 24th European CF Conference, Vienna, Austria.

Scottish Executive Health Department (2003). Protecting Patient Confidentiality: A Code of Practice for NHS Scotland Staff, Scottish Executive Health Department.

Sebaldt, R., G. Ioannidis, J. Adachi, et al. (1999). "36 month intermittent cyclical etidronate treatment in patients with established corticosteroid induced osteoporosis." Journal of Rheumatology **26**: 1545-1549.

Sernyak, M. J., R. Rosenheck, R. Desai, et al. (2001). "Impact of clozapine prescription on inpatient resource utilization." J Nerv Ment Dis **189**: 766-773.

Shah, P. I., A. Bush, G. J. Canny, et al. (1995a). "Recombinant human DNase I in cystic fibrosis patients with severe pulmonary disease: A short-term, double-blind study followed by six months open-label treatment." European Respiratory Journal **8**: 954-958.

Shah, P. L., S. Conway, S. F. Scott, et al. (2001). "A case-controlled study with dornase alfa to evaluate impact on disease progression over a 4-year period." Respiration **68**: 160-164.

Shah, P. L., S. F. Scott, H. J. Fuchs, et al. (1995b). "Medium term treatment of stable stage cystic fibrosis with recombinant human DNase I." Thorax **50**: 333-338.

Shah, P. L., S. F. Scott, D. M. Geddes, et al. (1995c). "Two years experience with recombinant human DNase I in the treatment of pulmonary disease in cystic fibrosis." Respiratory Medicine **89**: 499-502.

Shapiro, S. (1989). "The role of automated record linkage in the postmarketing surveillance of drug safety: a critique." Clinical Pharamcology and Therapeutics **46**: 371-386.

Simon, G., E. Wagner and M. Vonkorff (1994). "Cost-effectiveness comparisons using "real world" randomized trials: the case of new antidepressant drugs." Journal of Clinical Epidemiology **48**: 363-373.

Singh, G. (2001). "Arthritis, rheumatism and aging medical information system post-marketing surveillance program." The Journal of Rheumatology **28**: 1174-1179.

Skingle, S. and A. Crisp (1994). "Increased bone density in patients on steroids with etidronate." The Lancet **344**: 543-544.

Solomon, D. H., R. J. Glynn, R. Levin, et al. (2002). "Nonsteroidal anti-inflammatory drug use and acute myocardial infarction." Archives of Internal Medicine **162**: 1099-1104.

Sørensen, H., S. Sabroe and J. Olsen (1996). "A framework for evaluation of secondary data sources for epidemiological research." International Journal of Epidemiology **25**: 435-442.

Sørensen, H., T. Lash and K. Rothman (2006). "Beyond randomized controlled trials: A critical comparison of trials with nonrandomized studies." Hepatology **44**: 1075-1082.

Starmer, C. and K. Lee (1982). "A data-based approach to assessing clinical interventions in the setting of chronic disease." Cancer Treatment Reports **66**: 1077-1082.

Stegmayr, B., K. Asplund, K. Hulter-Åsberg, et al. (1999). "Stroke units in their natural habitat - can results of randomized trials be reproduced in routine clinical practice?" Stroke **30**: 709-714.

Stenberg, V., J. Fiechtner, J. Rice, et al. (1992). "Endocrine control of inflammation: Rheumatoid arthritis double-blind, crossover clinical trial." International Journal of Clinical and Pharmacological Research **12**: 11-18.

Sterne, J., M. Egger and G. Davey Smith (2001). "Investigating and dealing with publication and other bias". In: M. Egger, G. Davey Smith and D. Altman: Systematic Reviews in Health Care: Meta-Analysis in Context. London, BMJ Publishing Group: 189-208.

Stevens, A., J. Raftery and P Roderick (2005). "Can health technologies be assessed using routine data?" International Journal of Technology Assessment in Health Care 21: 96-103.

Stokes, D., M. W. Konstan, W. J. Morgan, et al. (1997). Risk factors for decline in pulmonary function in young and mild CF patients. European Working Group for Cystic Fibrosis Meeting, Davos, Switzerland.

Strobl, J., E. Cave and T. Walley (2000). "Data protection legislation: interpretation and barriers to research." British Medical Journal **321**: 890-892.

Strobl J., I. Enzer, A. Bagust, et al. (2003). "Using disease registries for pharmacoepidemiological research: a case study of data from a cystic fibrosis registry." Pharmacoepidemiology and Drug Safety **12**: 459-65.

Strom, B., K. Melmon and O. Miettenen (1985). "Post-marketing studies of drug efficacy: why?" The American Journal of Medicine **78**: 475-480.

Strom, B., O. Miettenen and K. Melmon (1984). "Post-marketing studies of drug efficacy: how?" The American Journal of Medicine **77**: 703-708.

Stürmer, T., M. Joshi, R. Glynn, J. Avorn, K. Rothman, S. Schneeweiss (2006). "A review of the application of propensity score methods yielded increasing use, advantages in specific

settings, but not substantially different estimates compared with conventional multivariable methods." Clinical Epidemiology **59**: 437-447.

Suri, R., M. C., B. Lees, et al. (2001). "Comparison of hypertonic saline and alternate-day or daily recombinant human deoxyribonuclease in children with cystic fibrosis: a randomised trial." The Lancet **358**: 1316-1321.

Sutton, A., K. Abrams, D. Jones, et al. (1998). "Systematic reviews of trials and other studies." Health Technology Assessment **2**.

Sutton, A., K. Abrams, D. Jones, et al. (2000). "Methods for meta-analysis in medical research." Wiley Series in Probability and Statistics, pp. 239-255. West Sussex, Wiley.

Talbot, S., A. K. Webb, G. Hambleton, et al. (1998). Co-operation between health care sectors - developing a novel tool to estimate compliance. unpublished conference abstract.

Tassie, J.-M., J. Gasnault, M. Bentata, et al. (1999). "Survival improvement of AIDS-related progressive mulitfocal leukoencephalopathy in the era of protease inhibitors." AIDS **13**: 1881-1887.

Temple, R. (1999). "Meta-analysis and epidemiologic studies in drug development and postmarketing surveillance." Journal of the American Medical Association **281**: 841-844.

Thompson, S. (2001). "Why and how sources of heterogeneity should be investigated". In: M. Egger, G. Davey Smith and D. Altman: Systematic Reviews in Health Care: Meta-Analysis in Context. London, BMJ Publishing Group: 157-175.

Tiefenbrunn, A. J., N. C. Chandra, W. J. French, et al. (1998). "Clinical experience with primary percutaneous transluminal coronary angioplasty compared with alteplase (recombinant tissue-type plasminogen activator) in patients with acute myocardial infarction: a report from the Second National Registry of Myocardial Infarction (NRMI-2)." J Am Coll Cardiol **31**: 1240-1245.

UKCC (1996). Guidelines for Professional Practice. London, United Kingdom Central Council for Nursing, Midwifery and Health Visiting.

UKCC (1998). Guidelines for Records and Record Keeping. London, United Kingdom Central Council for Nursing, Midwifery and Health Visiting.

van Gestel, A., R. Laan, C. Haagsma, et al. (1995). "Low-dose oral steroids as bridge therapy in rheumatoid arthritis patients starting with parenteral gold: a randomized double-blind placebo-controlled trial." British Journal of Rheumatology **34**: 347-351.

van Pinxteren, B., M. Numans, P. Bonis, et al. (2001). "Short-term treatment with proton pump inhibitors, H2-receptor antagonists and prokinetics for gastro-oesophageal reflux disease-like symptoms and endoscopy negative reflux disease (Cochrane Review)". In: The Cochrane Library. Oxford, Update Software.

van Schaardenburg, D., R. Valkema, B. Dijkmans, et al. (1995). "Prednisone for elderly-onset rheumatoid arthritis: Outcome and bone mass in comparison to treatment with chloroquine." Arthritis and Rheumatism **38**: 334-342.

van Staa, T. P., L. Abenhaim and C. Cooper (1998). "Use of cyclical etidronate and prevention of non-vertebral fractures." Br J Rheumatol **37**: 87-94.

von Elm, E., D. Altman, M. Egger, et al. (2007) "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. PLoS Med 4(10): e296 doi:10.1371/journal.pmed.0040296

Venables, T., R. Newland, A. Patel, et al. (1997). "Omeprazole 10 milligrams once daily, omeprazole 20 milligrams once daily, or ranitidine 150 milligrams twice daily, evaluated as initial therapy for the relief of symptoms of gastro-oesophageal reflux disease in general practice." Scandinavian Journal of Gastroenterology **32**: 965-973.

Walters, S. (2000). "Clinical epidemiology of cystic fibrosis". In: M. Hodson and D. Geddes: Cystic Fibrosis. London, Arnold: 2-12.

Weaver, W.D., R. Simes, A. Betriu, et al. (1997) "Comparison of primary coronary angioplasty and intravenous thrombolytic therapy for acute myocardial infarction: a quantitative review" JAMA **278**: 2093-2098.

Weintraub, J. A., S. C. Stearns, R. G. Rozier, et al. (2001). "Treatment outcomes and costs of dental sealants among children enrolled in Medicaid." American Journal of Public Health **91**: 1877-1881.

Wells, G., B. Shea, D. O'Connell, et al. (2003) "The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses." Department of Epidemiology and Commuunity Medicine, University of Ottawa, Canada. http://www.lri.ca/programs/ceu/oxford.htm (accessed on 14 July 2003).

Williams, H., J. Ward, J. Reading, et al. (1992). "Comparison of auranofin, methotrexate, and the combination of both in the treatment of rheumatoid arthritis: a controlled clinical trial." Arthritis and Rheumatism **35**: 259-269.

Williams, J., W. Cheung, D. Cohen, et al. (2003). "Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment." Health Technology Assessment **7**: 26.

Wilmott, R. W., R. S. Amin, A. A. Colin, et al. (1996). "Aerosolized recombinant human DNase in hospitalized cystic fibrosis patients with acute pulmonary exacerbations." American Journal of Respiratory & Critical Care Medicine **153**: 1914-1917.

Wilson, A., G. Grimshaw, R. Baker, et al. (1999). "Differentiating between audit and research: postal survey of health authorities' views." British Medical Journal **319**: 1235-.

Wizla-Derambure, N., L. Michaud, A. Sardet, et al. (1998). "Effects of rhDNAase on lung function and nutritional status in a cystic fibrosis pediatric population. [French]." Archives de Pediatrie **5**: 378-383.

Woodman, R. (2000). "Storage of human organs prompts three inquiries." British Medical Journal **320**: 77.

Worth, H., D. Stammen and E. Keck (1994). "Therapy of steroid-induced bone loss in adult asthmatics with calcium, vitamin D, and a diphosphonate." American Journal of Respiratory & Critical Care Medicine **150**: 394-397.

Wyshak, G., E. Burdick and F. Mosteller (1991). "Technology Assessment in the Connecticut Tumor Registry." International Journal of Technology Assessment in Health Care **7**: 129-133.

Zaza, S., L. Wright-de Agüero, P. Briss, et al. (2000). "Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services." American Journal of Preventive Medicine **18**(1S): 44-74.

Zelen, M. (1979). "A newdesign for randomized clinical trials." New England Journal of Medicine **300**: 1242-1245.

Zelen, M. (1981). "Alternatives to classic randomized trials." Surgical Clinics of North America **61**: 1425-1432.

Ziegelstein, R. C., J. M. Hilbe, W. J. French, et al. (2001). "Magnesium use in the treatment of acute myocardial infarction in the United States (observations from the Second National Registry of Myocardial Infarction)." American Journal of Cardiology **87**: 7-10.

# Appendices

# A Appendix A: Material relating to Dornase Alfa Case Study

- Literature review of CF and the effectiveness of dornase alfa
- Data entry forms used by the ERCF
- Formulae for predicted lung function used in the study
- Sample forms for sample DQR
- Further baseline data

# *Literature review: CF and the effectiveness of dornase alfa*

This section presents the background to the objectives of the dornase alfa study. An overview of the epidemiology of CF with a focus on prognostic factors relevant to this study forms the introduction to a more detailed review of the existing literature on dornase alfa, focussing on different methods used in estimating its efficacy and effectiveness.

## Epidemiology of CF

### *Incidence, prevalence and survival of CF in the UK*

Cystic fibrosis is an inherited disease, affecting an estimated one in 2,500 newborn babies in the UK (Dodge *et al.* 1997), but there are considerable national variations in reported incidence. The median life expectancy has increased significantly over the past decades due to comprehensive symptomatic treatment regimens. For a child born with CF today, the median life expectancy has been estimated to be around 40 years (Elborn *et al.* 1991). A recent estimate of the size of the UK's entire CF population in the year 2000 was 7,750, compared to 6,500 in 1992 (Dodge *et al.* 1997). This increase was experienced only in the adult population, with the child population having stabilised at some 4,500 children owing to now very low mortality rates in children. A more recent paper estimating the survival of adults with CF in the UK born between 1947 and 1967 points out that whilst the survival into later adulthood is rising, the mortality in those age groups does not appear to be improving (Lewis *et al.* 1999).

Factors associated with poor prognosis of CF are chiefly poor lung function and nutritional status, pancreatic in-sufficiency, but also a respiratory

colonisation with *Burkholderia cepacia,* and according to some studies also *Pseudomonas aeruginosa* (Walters 2000). There is also a broad consensus that patients treated by specialist centres enjoy better clinical outcomes. CF patients suffer a high prevalence of diabetes mellitus, especially in later life, and the presence of diabetes has also been associated with a considerably lower median survival (Koch *et al.* 1999). In most countries, the life expectancy of female patients is shorter than that of male patients, although it is not clear whether this is a gender-specific issue per se. Patients born in the UK in more recent years seem to show an equilibrium of age-specific mortality rates between gender (Lewis 2000).

### *Genotype*

CF is caused by abnormal function of the chloride channel CFTR. Numerous mutations of the gene encoding this CFTR have so far been identified (to date more than 800). A growing body of research is dedicated to the exploration of the relationship between genotype and phenotype in CF. The most common mutation in Caucasian populations seems to be DF508, which is reported to occur in 68% of CF alleles (Rosenstein 2000). A recent summary on this topic concludes that it is clear that defects in each copy of the relevant gene cause the CF phenotype, and that even among patients with the classic form of CF, variability can be observed. A second conclusion was that genotype correlates more closely with certain phenotypic features and that decreasing levels of CFTR function are associated with progressive involvement of more organ systems (Cutting 2000). Although some genotypes are associated with milder or less mild phenotype, prognosis cannot be determined by genotype analysis alone (Walters 2000).

### *Diagnosis*

An initial diagnosis of CF is based on the presence of typical clinical symptoms characteristic of the CF phenotype (e.g. chronic respiratory disease, malabsorption and intestinal obstructive disorders, including

meconium ileus, pancreatic insufficiency). The presence of these symptoms is then evaluated against family history or a positive neonatal screening test, and the evidence of genetic abnormality from elevated sweat chloride concentrations, or the identification of CF mutations. In a cross-sectional study of a US CF registry, 15.7% of patients had meconium ileus at birth recorded; at the time of diagnosis, respiratory symptoms were present in 44.6% of patients, failure to thrive or malnutrition in 35.6%, and steatorrhoea and malabsorption in 21.1% of patients (FitzSimmons 1994).

The US registry study found that in 1990 70% of patients were diagnosed during the first year of their lives, 80% by the age of four years, and 90% by the age of twelve years (FitzSimmons 1994). However, diagnosis in later life is always a possibility; this has been associated with milder clinical symptoms and their progression as well as a lower prevalence of *P. aeruginosa* infections (Gan *et al.* 1995). A Dutch study of 143 patients comparing those diagnosed before and after the age of 16 years found no patients homozygous for DF508 in the latter group (Gan *et al.* 1995).

One might assume that screening is associated with better prognosis, because treatment can commence earlier, but clinical advantages have proved difficult to demonstrate. A study based on data from the Epidemiologic Registry of Cystic Fibrosis (ERCF) compared patients who were diagnosed by neonatal screening (1,684 of 13,684 patients) with the rest of the registered population (Navarro *et al.* 1999). Whereas nutritional status of the under-13-year-olds was better in the group diagnosed by screening, an advantage in respiratory status of this group was only observed in the youngest patients under 6 years of age (Navarro *et al.* 1999).

### *Respiratory disease and lung function*

The main manifestation of CF is respiratory disease. The altered secretions of the respiratory tract lead to viscous mucus, which is

susceptible to infections with a variety of microorganisms. Persistent infections lead to chronic inflammation.

Forced Expiratory Flow in one second ($FEV_1$) is one of the most frequently used monitoring indicators and outcome measures in CF, together with Forced Vital Capacity (FVC). Both $FEV_1$ and FVC are usefully expressed in a standardised form as per cent of the value predicted for the age, sex, and height of a patient ("% of predicted"). $FEV_1$ is generally regarded as the most significant predictor of mortality amongst CF patients. A cohort study undertaken in the late 1970s and 1980s in Canada signalled that patients with an $FEV_1$ of less than 30% of predicted had a two-year mortality rate of over 50% (Kerem *et al.* 1992).

Once airways obstruction has sufficiently progressed, the lung function of CF patients deteriorates progressively; according to various cohort studies, this decline is approximately 3-5% of predicted $FEV_1$ per annum (Walters 2000). In terms of therapeutic effectiveness, any achieved delay, slow-down, or suspension of this decline is beneficial compared to continued decline. One of the main sources of data on annual lung-function decline to be expected in CF is a 4-year randomised trial of ibuprofen in CF. The placebo group showed an annual rate of decline in $FEV_1$ of -3.60±0.55% (mean slope±SE). However, the study only included patients with a baseline value of ≥60% of predicted $FEV_1$ and also excluded patients colonised with *B. cepacia* and other medical conditions relevant to the investigational drug (Konstan *et al.* 1995).

Braggion and colleagues (1997) retrospectively reviewed the decline in $FEV_1$ of a cohort of 57 selected patients between the ages of seven and twelve years. They too report an annual decline of -3.6±2.9% predicted $FEV_1$ (Braggion *et al.* 1997). Another Italian study followed 48 children aged 6-15 years for an average of 5.8±0.6 years and estimated a mean annual decline of -2.1 (range: -5 to 1.4) (Lucidi *et al.* 1997).

Some variation in the rates of decline has been put down to factors such as age or baseline $FEV_1$ value. Data from 366 patients on a Canadian database, who had at least one pulmonary function test recorded before they were 10 years of age, showed great variations in decline between subgroups of patients. The authors remark that average rates of decline were significantly (negatively) related to survival age. Women and patients with pancreatic insufficiency showed a steeper decline (Corey *et al.* 1997). A study of 5,313 ERCF patients with an observation period of ≥2 years identified an association of above average $FEV_1$ decline with diabetes, colonisation with *B. cepacia,* and severe malnutrition (Mastella *et al.* 1999).

The North-American sister database of the ERCF, the Epidemiological Survey of CF (ESCF) (Morgan *et al.* 1999) provided data for a longitudinal assessment whereby rapid and non-decliners in predicted $FEV_1$ values were defined amongst 6-12-year-old patients observed over 18 months, based on the upper and lower thirds of annual change in $FEV_1$. A body weight of $<10^{th}$ percentile, *P. aeruginosa* colonisation, and daily sputum production were independently associated with an increased risk of a rapid decline of $FEV_1$ (Stokes *et al.* 1997).

A cross-sectional analysis of ERCF data on patients aged 6 years and older indicates that an $FEV_1$ value of ≥10% less than predicted was associated with lower weight for age percentiles, pulmonary symptoms as well as colonisation with *P. aeruginosa* and *B. cepacia* (Navarro *et al.* 2001). Unsurprisingly, treatment-related variables such as the use of oral corticosteroids, non-steroidal anti-inflammatory drugs, dornase alfa, or oxygen were associated with lower $FEV_1$ values.

There is evidence that not only premature mortality but also lung function decline is reducing in the CF population; the Adult CF Unit in Manchester reported significant differences in the rate of lung function decline in their patients aged over 26 years, when comparing data from within one

decade (the periods 1985-6, 1989-90, and 1993-4); the latter periods showed the slowest decline (Abbott *et al.* 1995).

### *Anthropometrics*

Besides lung function, poor nutritional status is one of the main predictive factors in CF disease. In a cross-sectional study of some 3,000 patients from the majority of the UK's large CF specialist centres (i.e. nearly half of the CF population) during 1994/5, Morison *et al.* (1997) have shown that height and weight in the first 10 years of life is now at approximately 0.5 standard deviations (SD) below the population mean, despite poorer SD scores during the first year of life (possibly due to late detection of mild cases as well as some catch-up growth in the following years). It is also not clear what part low birth weight might play, if any. The weight of adults is still deviating considerably from the norm, particularly in men, which could be due to fewer severely affected women surviving to older age groups.

A US study of over 13,000 patients in 1993 indicated that mean and median height- and weight-for-age of CF children were at the $20^{th}$ and $30^{th}$ percentiles based on the National Center for Health Statistics/Centers for Disease Control growth reference. Twenty per cent of the children were below the $5^{th}$ percentile (Lai *et al.* 1998). A study based on ERCF data indicates that adolescents with CF maintain a constant disadvantage of about one standard deviation from the normal growth curve (Navarro *et al.* 2000). The youngest age group (<6 years) and adults of 18 years or older seem to show a slight catch-up growth; in 6-12-year-olds the opposite was observed. However, it is argued that given the present treatment possibilities, nutritional state and growth in most CF patients should be normal (Rosenstein 2000).

### *Microbiological colonisation*

Recurrent and chronic infections of the viscous sputum are amongst the main complications of CF. The organism most frequently found in sputum

cultures, particularly in the UK, is *Pseudomonas aeruginosa*, which tends to be more prevalent in older age groups. Similarly, *Burkholderia cepacia* is more likely to be found in older patients, albeit that its prevalence is much lower. For both organisms prevalence varies between countries and centres. *B.cepacia* is often seen as a factor associated with worse prognosis (Walters 2000). According to a matched case-control study involving 125 patients, *B. cepacia* infection per se is not necessarily associated with poor pulmonary status, but is associated with poor long-term survival (Frangolias *et al.* 1999). Similarly, data from the ERCF indicates that colonisation with *B. cepacia* was associated with shortened survival (Koch *et al.* 1999) and lower average change per year in weight-for-age Z-score (Navarro *et al.* 2000).

### *Exacerbation*

Pulmonary exacerbations of CF are generally indicated by a raised temperature, cough, increased sputum production, weight loss, decreased lung function or similar symptoms indicating an inflammatory response to an infection. A one-year retrospective review of 299 adult patients in 3 CF centres in the UK estimated the annual rate of exacerbation episodes as 3.54 per patient; this varied between centres, presumably also partly due to differences in defining diagnostic criteria for exacerbations (Bilton *et al.* 1998).

### *Diabetes*

Diabetes is particularly common in CF and rising with age. Of 114 CF patients aged 30-59 years and registered with the French registry of CF (male/female: 1.6), 12% had insulin-dependent diabetes mellitus (Badet *et al.* 2001). Similarly, the prevalence of diabetes mellitus in the Danish CF population has been reported to be 14.7%, (Lang *et al.* 1994) and the mean age at diagnosis 21 years (Lang *et al.* 1995). The presence of diabetes has been strongly linked to poorer lung function across all age groups as well as poor nutritional status (Koch *et al.* 2000).

## Efficacy and effectiveness of dornase alfa

The following sections review the evidence on dornase alfa depending on the type of study design used (i.e. randomised and non-randomised studies). For this review, the bibliographic databases Medline and Embase were searched for "(DORNASE ALFA OR DORNASE OR DEOXYRIBONUCLEASE OR PULMOZYME) AND CYSTIC FIBROSIS". Original reports of studies evaluating the drug were included in the review. In addition, abstracts of the European CF Conferences since the licensing of dornase alfa (1994) have been searched.

### *Evidence from randomised trials*

A recent Cochrane review addressed the question of whether dornase alfa was associated with reduced mortality and morbidity compared to placebo (Kearney & Wallis 2003). The authors conclude that the available studies were of insufficient duration to assess the impact on mortality or whether the identified beneficial effect on lung function is sustained in the long-term. Hence, there was no evidence that dornase alfa affects the natural history of lung disease in CF patients in the long-term or that long-term use improves life expectancy. All seven trials included in the Cochrane review are represented here, together with more recently published studies as well as some studies excluded by the Cochrane review (studies not included there were: (Hubbard *et al.* 1992; Robinson *et al.* 2000; Nasr *et al.* 2001; Quan *et al.* 2001).

Aitken and colleagues (1992) reported an early non-randomised trial to assess the safety of dornase alfa on 12 healthy subjects and 14 CF patients with an FVC of >40% predicted and no recent exacerbations. After a 2-week treatment period and later re-challenge with a single dose the authors concluded that dornase alfa appeared safe in both normal and CF subjects. Another early and similarly small study (Hubbard *et al.* 1992) confirms these safety results.

Of four further short trials (Ramsey *et al.* 1993; Ranasinha *et al.* 1993; Laube *et al.* 1996; Wilmott *et al.* 1996), three (Ramsey *et al.* 1993; Ranasinha *et al.* 1993; Laube *et al.* 1996) showed significant improvements in lung function over placebo (see Table A.1). The phase II study by Ranasinha and colleagues (1993) was followed by a 6-month open-label extension of 59 patients. This showed a further improvement during the first month and than stabilised at a mean increase of 6.2% $FEV_1$ (CI: 4.6-7.8%) and 7.2% FVC (CI: 5.1-10.2%) from baseline (Shah *et al.* 1995b). Two weeks after the therapy was discontinued, $FEV_1$ and FVC fell 5.7% and 5.9% below the baseline respectively. A further 18-month open-label extension involving 52 patients was initiated using 2.5mg dornase alfa once daily (rather than twice daily as before). During these 18 months, $FEV_1$ and FVC increased again by a mean 8% and 1.2% respectively from the new baseline (Shah *et al.* 1995c). This improvement was accompanied by a gradual weight gain also. A recent 7-day RCT, with ciliary and cough clearance as its main outcome of interest, observed a significant increase from baseline, but the difference between the improvement of the placebo and treatment group was only significant for FVC, but not $FEV_1$ as outcome measure (Robinson *et al.* 2000).

There are so far only two RCTs of severely affected patients (FVC <40%), for whom the drug is currently not licensed. The first showed an improvement in $FEV_1$ in the dornase alfa group of 9.4% from baseline, compared with 2.1% in the placebo group ($p<0.001$) after three months (McCoy *et al.* 1996). The randomised phase of the second study was only 2 weeks and did not show a significant difference between the two groups in mean percentage change from baseline in per cent of predicted $FEV_1$ or FVC ($FEV_1$: 1.4% [SEM 2.1] in the dornase alfa group, and 4.2% [SEM 2.2] in the placebo group; FVC: 8.8% [SEM 3.6] in the dornase alfa group, and 13.7% [SEM 3.8] in the placebo group) (Shah *et al.* 1995a). Sixty-four patients entered a 6-month open-label extension of this trial, but only 38 completed the study; the overall mean improvement was 9% of predicted

$FEV_1$ and 18% of predicted FVC. The efficacy of dornase alfa in severe patients is thus still under question (Newman *et al.* 2001).

Most studies have included patients from age five or more upwards. Nasr and colleagues (2001) performed a placebo-controlled randomised trial of dornase alfa over 100 days on 12 children under 5 years of age. The outcome measures used were high-resolution computer tomography (HRCT) of the chest and chest x-ray (CXR), since spirometry measures are not reliable in small children. Two scorers rated the scans and x-rays and average scores between them were used for the analysis. The HRCT but not the CXR scores showed a significant improvement in the mean change from baseline between the placebo and treatment groups.

Until recently, the most important RCT was a multi-centre 6-month phase-III trial, involving 968 patients (Fuchs *et al.* 1994). Main outcome variables were the occurrence of exacerbations of respiratory symptoms requiring parenteral antibiotics, and pulmonary function. The age-adjusted risk of a defined exacerbation requiring parenteral antibiotic treatment was reduced by 28% (relative risk, 0.72; 95%CI: 0.52 - 0.98; p=0.04) in patients receiving 2.5mg dornase alfa once daily and 37% (relative risk 0.63; 95%CI: 0.46 - 0.87; p<0.01) in those receiving the same dose twice daily, compared to a placebo group (for all exacerbations - including those not meeting the study's criteria - see Table A.1). Over the duration of the study, $FEV_1$ improved on average by 5.8% (SD: 0.7) and 5.6% (SD: 0.7) in the group treated with 2.5mg dornase alfa once and twice daily respectively (p<0.01 compared with placebo). Compared with the placebo group, those receiving 2.5mg dornase alfa once daily spent 1.3 fewer days in hospital (p=0.06), those receiving the dose twice daily, 1.0 fewer days (p=0.5). The first treatment group spent 2.7 fewer days receiving parenteral antibiotics (p=0.05), the second 2.2 fewer days (p=0.13), compared with the placebo group.

In December 2001, results of a 2-year RCT funded by Genentech, Inc, and Hofmann-La Roche, Ltd. were published (Quan *et al.* 2001). Patients were of a mean age of 8.4 years and had mean $FEV_1$ of 95% of predicted at baseline (FVC: 102%). After an initial improvement following initiation of dornase alfa treatment, mean $FEV_1$ values of the treatment group returned to baseline after 96 weeks. The advantage over the control group was a mean 3.2% predicted $FEV_1$ (SD: 1.2; p=0.006), however, at one year follow up, this was not significantly different. The risk of respiratory tract exacerbations was reduced by 34%, which was not statistically significant (95%CI: 0.44-1.00; p=0.48).

## Table A.1: Overview of randomised controlled trials assessing dornase alfa

| Reference | Sample characteristics[a)] | Duration of treatment | $FEV_1$ (Mean change in treatment group) | FVC (Mean change in treatment group) | Exacerbations | Comments and daily dose |
|---|---|---|---|---|---|---|
| (Fuchs *et al.* 1994) | 968 patients<br>Age: 5 years +,<br>FVC: 33-100%pred. | 24 weeks | 5.8% (SD: 0.7) (1x daily)<br>5.6% (SD: 0.7) (2x daily)<br>(Significantly better than placebo: p<0.01) | 3.8% (SD: 0.6) (1x daily)<br>3.0% (SD: 0.6) (2x daily)<br>(Significantly better than placebo: p<0.01) | Reduction in age-adjusted risk of all exacerbations by 31% (p<.01) (1x daily), 32% (p<.01) (2x daily) | Dose: 2.5mg (1x daily), 5mg (=2.5mg 2xdaily) |
| (Hubbard *et al.* 1992) | 16 patients<br>Mean age: 27±1 years<br>$FEV_1$:>1.1 litres | 6 days | 0.26 litres (SE: 0.07)<br>(Significantly better than placebo: p<0.01) | 0.35 litres (SE: 0.10)<br>(Significantly better than placebo: p<0.01) | | Crossover trial.<br>Dose: 20mg |
| (Laube *et al.* 1996) | 20 patients<br>Age: >18 years<br>FVC: 35-75%pred. | 6 days | 9.4% (SE: 3.5) pred.<br>(Significantly better than placebo: p<0.05) | 12.7% (SE: 2.6) pred.<br>(Significantly better than placebo: p<0.05) | | Dose: 5mg |
| (McCoy *et al.* 1996) | 320 patients<br>Age: 7-57 years<br>FVC: <40%pred. | 12 weeks | 9.4% (CI: 6.81-11.96)<br>(Significantly better than placebo: p<0.001) | 12.4% (CI: 9.46-15.33)<br>(Significantly better than placebo: p<0.01) | No difference in dyspnoea score, days on intravenous antibiotics, or days in hospital | Dose: 2.5mg |
| (Nasr *et al.* 2001) | 12 patients<br>Age: <5 years | 100 days | This study assessed efficacy by high-resolution computerised tomography (HRCT) of the chest and chest x-ray (CXR); significant difference in mean change from baseline between placebo and treatment group on HRCT (p=0.02), but not CXR (p=0.16). | | | |
| (Quan *et al.* 2001) | 474 patients<br>Mean age: 8.4 yrs.<br>Mean FEV1: 95%pred. | 96 weeks | 0.04% (SE: 0.8) pred.<br>(Significantly better than placebo: p=0.006) | -2.2% (SE: 0.7) pred.<br>(Not significantly different from placebo) | Dornase alfa reduced risk of respiratory tract exacerbations by 34% (p=0.48) | Dose: 2.5mg |
| (Ramsey *et al.* 1993) | 181 patients<br>Age: 8-65 years<br>FVC: ≥ 40%pred. | 10 days | 9.9% (SE: 2.0) (on 1.2mg)<br>13.8% (SE: 2.0) (on 5mg)<br>14.5% (SE: 2.1) (on 20mg)<br>(Significantly better than placebo) | 9.9% (SE: 2.0) (on 1.2mg)<br>11.8% (SE: 1.9) (on 5mg)<br>9.6% (SE: 1.5) (on 20mg)<br>(Significantly better than placebo) | | Dose:<br>1.2mg, 5mg, 20mg |
| (Ranasinha *et al.* 1993) | 71 patients<br>Age: 16-55 years<br>FVC: >40%pred. | 10 days | 13.3%<br>(Significantly better than placebo: p<0.001) | 7.2%<br>(Not significantly different from placebo) | | Dose: 5mg |

| Reference | Sample characteristics[a)] | Duration of treatment | $FEV_1$ (Mean change in treatment group) | FVC (Mean change in treatment group) | Exacerbations | Comments and daily dose |
|---|---|---|---|---|---|---|
| (Robinson *et al.* 2000) | 15 patients<br>Age: >18 years<br>FVC: 35-116%pred. | 7 days | 7.5%<br>(Not significantly different from placebo phase) | 5.4%<br>(Significantly better than placebo phase) | | Cross-over study, which assessed cough and mucociliary clearance primarily |
| (Shah *et al.* 1995a) | 70 patients<br>Age: 5-48 years<br>FVC: <40 %pred. | 2 weeks | 1.4% (SE: 2.1) pred.<br>(Not significantly different from placebo) | 8.8% (SE: 3.6) pred.<br>(Not significantly different from placebo) | | Dose: 5mg.<br>Further improvements on 6-month open-label extension. |
| (Wilmott *et al.* 1996) | 80 patients with acute respiratory exacerbation<br>Age: ≥ 5 years<br>FVC: ≥ 35%pred. | 2 weeks | Improvement, but not significantly different from placebo group | Improvement, but not significantly different from placebo group | | Dose: 5mg |

[a)] Patient numbers represent number of recruited patients.

## Table A.2: Non-randomised studies with control groups assessing dornase alfa

| Reference | Treatment group | Study design and comparisons | Duration of treatment | $FEV_1$ | FVC | Exacerbations | Comments |
|---|---|---|---|---|---|---|---|
| (Furuya *et al.* 2001) | 21 patients<br>Mean age: 11.1±.5 years<br>FVC: >30%pred. | 3 months treatment followed by 3 months placebo for same group | 3 months | Baseline: 73.5±4.8%pred.<br>After treatment: 88.6±5.9%pred.<br>After placebo: 74.7±4.6%pred. | Baseline: 83.6±4.4%pred.<br>After treatment: 95.5±5.0%pred.<br>After placebo: 85.5±4.3%pred. | No significant difference in number of patients with exacerbations (19% and 29% of patients on dornase alfa and placebo respectively) | Improvement following treatment statistically significant |
| (Hodson *et al.* 1999) | 431 patients with $FEV_1$ values 2 years apart and dornase alfa treatment of 2 years minimum | Data from the ERCF; Comparison group: 2,125 patients never having received dornase alfa | 2 years | Mean change from baseline: -0.7% (control group: -2.2%) | Not reported | | Younger patients and patients with milder disease benefited more |
| (Hodson *et al.* 1999) | 493 patients with treatment of 1 year minimum after 1 year without dornase alfa | Data from the ERCF; Comparison group: patients never having received dornase alfa | 1 year | | | Frequency of exacerbations changed by -0.2 (+0.1 in control group) vis-à-vis previous untreated year | Patients aged 6 to <13 appeared to benefit most (reduction of 4 exacerbations / 10 treated patients / year) |
| (Hodson *et al.* 2003) | 374 patients enrolled on ERCF prior to starting treatment | Data from the ERCF; Comparison group: patients never having received dornase alfa | 2 years | Baseline: 60.2± (SD:23.1)<br>After 1 year: 2.5% change; after 2 years: 0.3% change | Not reported | Exacerbation frequency compared to control group (year 1 to year 2):<br>-0.25 (CE: -0.12 to -0.39) | Younger patients benefited more |
| (Eisenberg *et al.* 1997) | 184 patients,<br>Age: 8-65 years | Open-label study of 14 days treatment followed by 14 days without treatment for 6 treatment cycles | 14 days x 6 (interrupted with 6 14-day washout periods) | Baseline (mean±SE): 59.8±1.9%pred.<br>Mean % change between 7.6-10.5% by day 15 of each cycle;<br>During each washout phase FEV1 returned to baseline | Baseline (mean±SE): 79.0±1.6%pred.<br>Mean % change between 4.1-6.1% by day 15 of each cycle;<br>During each washout phase FVC returned to baseline | | Daily dose: 20mg |

| Reference | Treatment group | Study design and comparisons | Duration of treatment | $FEV_1$ | FVC | Exacerbations | Comments |
|---|---|---|---|---|---|---|---|
| (Johnson *et al.* 1999a) | 283 patients on treatment, plus 191 not on treatment continuously (involved in intention-to-treat analysis)<br>Age: ≥6 years<br>$FEV_1$: ≥40%pred. | Data from ECSF; Comparison group: 2382 patients who had never received dornase alfa | 329 (SD: 45) days on average in treatment group; 139 (SD: 129) days in additional group of 191 patients involved in intention-to-treat analysis; | Mean $FEV_1$ improved by 3.9% of predicted; untreated group: -1.6% (benefit: 4.3% (SE=0.9; p<0.0001) | Not reported | Not reported | Baseline characteristics differed considerably (treated patients had lower pulmonary function, more colonisation and more exacerbations within last 6 months) |
| (Milla 1998) | 190 patients<br>(At initiation of dornase alfa:)<br>Age: 5-53 years<br>$FEV_1$: mean 82.89%pred (SD=28.65) | Retrospective study comparing data on lung function from two years prior and two years post initiation of dornase alfa (same patients) | 2 years | Slope: -2.62% predicted/year (significantly better than in preceding 2 years without treatment - p<0.0001) | Not reported | 0.52 and 0.56 admissions for exacerbations / patient / year in periods before and after dornase alfa prescription respectively (not significantly different) | |
| (Shah *et al.* 2001) | 38 patients<br>Age: >16 years<br>FVC: >40%pred. | 38 matched (for age, sex, pulm. function) controls from two other centres | 4 years | Slope of median change: -0.75 (-2.19 in control group) p=0.076 | Not reported | Median 3.13 (CI: 1.25-4.25) exacerbations in control group; 1.25 (CI: 0.63-3.0) in dornase alfa group (p=0.035) | Median number of days on iv antibiotics also significantly different between the groups (p=0.034) |

### *Evidence from non-randomised studies with control groups*

An overview of non-randomised studies with control groups is presented in Table A.2. The possibly longest available follow-up period is documented in a matched case-controlled study lasting four years (Shah *et al.* 2001). Over a 4-year period, the difference in the rate of decline of the treatment and control groups did not reach statistical significance. However, if data from the 2 years preceding dornase alfa treatment are included, the difference in the rate of decline is significant (decrease of the slope from -1.68 to -0.57 in treatment group, vs. increase from -0.76 to -2.19 in control group; p=0.002). Of the non-randomised controlled studies, this seems to be the only one having demonstrated a significantly improved exacerbation rate in the treatment group. There was a significant difference in the number of infective episodes experienced during the 4-year period (median 3.13 [IQR: 1.25-4.25] in the control group, vs. 1.25 [0.6-3.0] in the treatment group (p=0.035)). The median numbers of days of intravenous antibiotic treatment differed to a similar extent.

Johnson *et al.* (1999a) have analysed ESCF data to assess the effectiveness of dornase alfa using a multiple regression analysis to account for 49 potentially biasing factors. The follow-up period was only one year, and the study does not report any analysis of deceased patients' records or give any information on survival. The estimated benefit over the control group from using dornase alfa was 4.3% of predicted $FEV_1$ (3.2% on the intention-to-treat analysis which included patients who have been on dornase alfa only for part of the observational period). The crude mean values of $FEV_1$ for treated patients improved by 3.9% of predicted $FEV_1$ (95%CI: 2.14 - 5.67) compared with a decline of 1.6% (95%CI: 1.01 - 2.19) in the untreated cohort. This study also shed some light on the differences in populations on and off dornase alfa. Those treated had lower pulmonary function, more bacterial colonisation, and more exacerbations at baseline.

A conference abstract presented in 1999 (Hodson *et al.* 1999) reported an analysis of ERCF data: $FEV_1$ values of 431 patients who had received dornase alfa for 2 years were compared with the values for 2,125 patients who had not received the drug. Patients on dornase alfa for a shorter time were excluded from the study. Overall mean change in $FEV_1$ was -0.7% in the treated group, compared with -2.2% in the patients not treated, less encouraging than the findings of the North-American study of ESCF data (Johnson *et al.* 1999a) which used half the follow-up period. On the ERCF study, a comparison of untreated patients with patients who had one year without treatment followed by one year of treatment showed significantly fewer exacerbations in the treated group; males, patients between 6 and 12 years of age, and those without bacterial colonisation showed better responses. No lung function data are reported for this before-after comparison.

In 2003 - after the completion of the Dornase Alfa Case Study - Hodson and colleagues published another study of dornase alfa based on ERCF data (Hodson et al. 2003), comparing patients after 2 years of treatment with an untreated comparison group. The analysis was a simple comparison of the two groups without statistical control of known baseline differences in age and lung function. Treated patients had improved lung function after 1 year (by 2.5%), but not after 2 years (0.3%), compared to deterioration of -1.1 and 2.3% predicted after 1 and 2 years in the untreated group. The frequency of exacerbations had improved by -0.25 (CI: -0.12 to -0.39) in the treatment compared to the untreated group (comparison between year 1 and 2).

A retrospective study using data on lung function from the two years prior and two years post initiation of dornase alfa treatment found a mild significantly improved decline with treatment compared to the period without treatment (difference in slopes -2.08, CI: -1.13 - -3.03, $p<0.0001$) (Milla 1998). The author explained that the level of pulmonary function in his patient group has

been considerably better than that of previously investigated patients, and that the subjects belonged to one single practice where an intensive treatment regimen already focused heavily on mucus clearance.

A French study on 55 patients aged 3 to 27 years reported no change in $FEV_1$ during a year of follow-up (from 52.1% [SD: 23.5] to 53.5% [SD: 19.5]), whereas there was significant deterioration during the preceding year in which patients were treated with the standard aerosolised sodium 2 mercapto ethane sulfonate and oral ambroxol (57.1% [SD: 15.8] to 52.1% [SD: 23.5]; $p<0.05$) (Bertolo-Houriez *et al.* 1997).

A small German open-label study of dornase alfa observed 12 patients with mild to moderate lung disease (FVC: 45-107% of predicted) for 18 months, including a 3-month treatment pause after 12 months. The overall positive response after the first 12 months was remarkable and also continued to rise throughout the year, reaching a mean difference from baseline of 21.8% (CI: -25.5 to 76.1) in $FEV_1$ (Heuckmann *et al.* 1999). After the 3-month treatment interruption, the difference from baseline had fallen to 8.4% (CI: -27.5 - 33.9), but rose to 23.4% (CI: -13.7 - 52.6) after a further 3 months treatment. FVC results mirrored those observed on $FEV_1$, albeit at lower levels. Whereas all 12 patients improved in the first year, the response shown during that period interestingly was repeatable only in 5 of 12 patients (difference <5%) during the second treatment period (Heuckmann *et al.* 1999).

### *Follow-up studies without control groups*

There is one large multi-national study of 974 patients on dornase alfa (baseline FVC before initiation: 40-70% of predicted; age 5 years or over) (Harms *et al.* 1998a) which reported a mean improvement from baseline of 10.5% in $FEV_1$ and 7.2% in FVC after 12 weeks, a result which compares very well with the evidence available from many trials.

Small "open label" studies without control groups are frequently reported, particularly at conferences. Most of these studies include few patients, lack comparison groups, and are of short duration (typically less than 4 weeks). However, some of these studies have followed patients for a longer period and are mentioned here. It has to be said that in many cases, the reports are only available in abstract form and therefore allow only a very limited critical appraisal. Publication bias may be significant.

Amarales and colleagues (1998) report a 12-month follow-up of 18 patients aged 8-24 who experienced a mean improvement of +6.7% in FVC ($p<0.037$), and +8.4% $FEV_1$ ($p<0.025$). However, typically the response varied greatly, and on both those outcome measures 4 patients deteriorated. A study following 63 Czech patients over one year also reported a statistically significant improvement over the baseline values of $FEV_1$ ($p<0.05$) (Bartošová *et al.* 1999). A retrospective review of data on 65 children during their first year of dornase alfa therapy reported a median increase of 11.1% (CI: 0-18.8) in $FEV_1$ and 5.6% (CI: 0-17) in FVC after 9 months with substantial variability of responses between patients (Davies *et al.* 1997).

A retrospective review was undertaken on 71 patients using dornase alfa in Leeds, some of whom could be followed for up to 6 years. Overall the per cent of predicted $FEV_1$ values of all patients declined through every year (0.7% from baseline after one year, 6.3% after 2 years and then by approximately 3% per year to 5 years) (Ratnalingam *et al.* 2001). A sub-analysis comparing responders (those with at least 5% increase in $FEV_1$ or $FEF_{25-75}$ after 1 month) to non-responders found improved $FEV_1$ values in the responders for the first two years, compared with a fall in values in non-responders.

Significant differences from baseline in $FEV_1$ or FVC after one year in mild to moderate disease have been found by several authors (Nousia-Arvanitakis *et al.* 1997; Nowakowska *et al.* 1997). Other studies found a similar response

after two (Wizla-Derambure *et al.* 1998; Ratnalingam *et al.* 2001; Santos *et al.* 2001) or three (Sands *et al.* 2000) years. In two of these studies (Ratnalingam *et al.* 2001; Santos *et al.* 2001), the improvement was not significant after three or more years.

### *Comparisons with hypertonic saline*

A recent open cross-over trial of 12 weeks duration examined whether daily administration of dornase alfa was equally effective as alternate-day treatment, or treatment with hypertonic saline in 48 children (Suri *et al.* 2001). The increase in $FEV_1$ over 12 weeks was comparable for the two dornase alfa treatment groups: 16% (SD 25%) and 14% (SD 22%) for daily and alternate-day dornase respectively, the difference being not statistically significant (2%; 95%CI: -4 to 9%; p=0.01). However, $FEV_1$ increased only by a mean 3% (SD 21%) following treatment with hypertonic saline. Daily dornase treatment resulted in an 8% (CI: 2 to 14) greater increase in $FEV_1$ than did saline (p=0.01). Most patients (83%) had been dornase alfa users before the trial, and the response varied considerably between individual patients. Previous short-term and pilot studies had shown larger increases of $FEV_1$ with hypertonic saline: 15% increase after 2 weeks in a study by Eng (1996), and 8% after a 3-week pilot study, which used a larger volume of saline (Ballmann & Hardt 2002). Suri and colleagues (2001) suggest that the initial good response to hypertonic saline may not be sustained over time.

## Cost-effectiveness of dornase alfa

The study of Fuchs *et al.* (1994) has been used to assess the effects of dornase alfa on the cost of treating respiratory tract infections (RTIs). Over 24 weeks, the treatment of RTIs for patients receiving dornase alfa was estimated to be $814-1,682 less, but this did not take into account the cost of dornase alfa itself. The economic evaluations based on Fuchs' data concluded that the

cost of the drug was only offset to about 15-35% by reductions in pulmonary exacerbations and hospitalisations (Oster *et al.* 1995; Menzin *et al.* 1996).

In a similar study - also based on Fuchs' data - Schulenburg and co-workers (1995) estimated the cost-effectiveness of dornase alfa in a German context and with a health insurance perspective. Again, the drug costs were omitted from the analysis. The study reported a difference in net direct total costs of 1,970 German Marks after 6 months treatment (roughly £6,000). Treatment costs are approximately twice that high.

A Canadian Health Technology Assessment of dornase alfa also dealt with its clinical and economic impacts (Perras & Otten 1996). Based on a record review of a very small number of patients over a period of 12 and 15.9 months, the authors tentatively note that the drug may become more cost-effective over time, but such a conclusion was premature given the lack of information on long-term effectiveness of dornase alfa. The outcome of interest was rate of hospitalisation. Costs included those associated with hospitalisation, antibiotic use, and, of course, dornase alfa use.

The most recent report by the Wessex Institute on the topic has modelled the possible long-term benefits of dornase alfa, for want of evidence of long-term effectiveness (Christopher *et al.* 1999). An attempt has been made to estimate costs per life year gained. The authors model $FEV_1$ decline based on findings from Fuchs *et al.* (1994) and studies on the prediction of lung function decline (Kerem *et al.* 1992; Konstan *et al.* 1995). Their resulting cost-effectiveness calculations conclude that if all patients were treated throughout their lifetime after their $FEV_1$ fell below a typical level (a duration of treatment of 30 years was assumed), two additional life years might be gained at a discounted cost of £52,550 each (range between £25,000 and £57,000, depending on sensitivity analysis); if therapy was restricted to patients with an $FEV_1$ of 70% of predicted or less who responded to treatment,

seven life years might be gained at a cost of £16,110 each (range: £18,000 - £36,600) (it was assumed that 30% of patients tried on dornase alfa had a mean improvement of 20% in $FEV_1$; a duration of treatment of 37 years was assumed).

McIntyre (1999) (of Roche Products Ltd) reports the results of her model as giving two extra years of life at a cost per life year gained between £34,854 and £45,234. Her assumptions differ inasmuch as she does not allow for the selection of responders for treatment but rather assumes everyone to be treated below 70% of predicted $FEV_1$, and she takes account of the costs offset by dornase alfa (Oster *et al.* 1995).

Johnson *et al.* (1999b) have estimated the cost of treatment for Alberta (Canada) patients with CF under conventional (not randomised) conditions using individual-level data, and determined the impact of dornase alfa therapy - amongst other factors - on cost of care, using multivariate linear regression analyses. Patient and resource use data from the ESCF and cost data from US sources were used. Known predictors of morbidity and mortality in CF tended to determine use and cost of care, and severity of disease was an important determinant of cost, independent of dornase alfa. However, having received dornase alfa significantly contributed to higher costs. It accounted for 44% of estimated overall care costs of the 303 patients included in the sample.

A trial by Suri and colleagues suggests that alternate day dornase alfa use is equally effective as daily administration (Suri *et al.* 2001). The mean incremental cost of using daily rather than alternate treatment was £513 (CI: £-546 to £1,510) over a 12-week period; daily dornase alfa use rather than hypertonic saline incurred a mean incremental cost increase of £1,409 (CI: £440 to £2,318). It was mentioned earlier that there was no difference between daily and alternate-day dornase alfa use over the trial period (2% difference in mean increase of $FEV_1$ (CI: -4 - 9, p=0.55).

## Further observations regarding the use and effectiveness of dornase alfa

There are several issues raised by efficacy and effectiveness evaluations of different types; these issues, which mainly relate to treatment response and its observed variations, are discussed here in turn.

### *Varying response*

One of the most elusive questions regarding dornase alfa use concerns the significant proportion of patients who do not experience a favourable response to the therapy. Significant variations in the quality of the response has been observed and emphasised by many authors.

A Spanish study examined medical records of 199 patients (mean age 14.5 years, mean $FEV_1$ 54.1%, mean FVC 65.5%) using dornase alfa with a minimum follow-up period of 1 year (Cobos *et al.* 2000). The authors drew particular attention to the large inter-individual variability in response to treatment, with benefits being doubtful in some 50% of patients included in their study, whose $FEV_1$ had fallen over a 2-year period. Over the same period, 34% had improved 10% or more over their baseline values. Overall, a mean change in $FEV_1$ of 3.3% (CI: -1.1 to 7.6) and 5.1% (CI: -0.7 to 10.9) was reported after one and two years respectively. The medium-term response was correlated with the early response shown during the first 3 months. Similarly, Blau *et al.* (1997) reported that only 30-50% of patients with an FVC between 40-70% predicted had experienced significant improvements in their pulmonary function after 3 months. After 6 months, 11% of patients had experienced a fall in FVC of more than 20% from the baseline. Overall, both FVC and $FEV_1$ had improved at the 1-month assessment, but fallen again thereafter.

In the retrospective study by Milla (1998), who compared individual lung function changes from the 2-year period before and 2-year period after

initiation of dornase alfa treatment, over 60% of the 190 patients showed no change in their $FEV_1$ trends, nearly 30% experienced a decline and less than 10% experienced an improvement. Milla stresses the importance of assessing the benefit of treatment on an individual basis, also allowing for adverse responses.

No study seems to have found a way to predict response to therapy from any baseline values, including pulmonary function. Response after 6, 9 and 12 months correlated well with the response found after 3 months in a retrospective study of 65 children using dornase alfa (Davies *et al.* 1997), a finding similar to that of another study (Davies *et al.* 1997; Wizla-Derambure *et al.* 1998). Most centres therefore operate initiation protocols, monitoring the response in each eligible patient.

However, there is also an indication that the same individual may show a different response at different times. A very small German study on 12 patients reported that only in 5 patients could the response observed during one year of treatment be repeated (difference <5%) after a 3-month treatment interruption and further 3 months treatment (Heuckmann *et al.* 1999). Finally, it has been suggested that poor inhalation technique could be responsible for a poor response and that a more intelligent aerosol delivery system which can compensate for poor technique may benefit patients who previously have not shown an adequate response (<10% increase in $FEV_1$ during a period of 2 weeks) to dornase alfa treatment (Scott *et al.* 2001).

### *Severity at baseline*

There are indications that patients with poorer lung function at baseline show less improvement following treatment with dornase alfa than patients with mild or moderate disease (Cimino *et al.* 1997). However, a small Spanish study saw less improvement after 21 months in the subgroup with >70% predicted FVC at baseline compared with a group who had baseline values

between 40 and 70% (Dapena *et al.* 1997). A French study reports similar findings after a 2-year follow-up (Deneuville *et al.* 1997), and another small 1-month open-label study also observed that initial lung function was significantly less in responders (13 patients who showed a 10% or greater increase in $FEV_1$ from baseline out of a total sample of 20 patients) than non-responders (Henry *et al.* 1998). Observations from the ERCF signal that young patients and those with mild disease may benefit most from dornase alfa treatment (Hodson *et al.* 1999).

### *Deterioration after treatment*

Some authors have reported significant decreases in lung function several days after treatment, despite initial significant improvements (Shah *et al.* 1995b; Amelina *et al.* 1999). In the early short RCTs, significantly improved spirometry values after 10 days had deteriorated back to baseline values one month after cessation of therapy (Ramsey *et al.* 1993; Ranasinha *et al.* 1993).

Furuya and colleagues (2001) observed patients during a 3-month treatment period followed by a 3-month period on placebo. The significant gain in lung function during the first three months was followed by a return to baseline values at the end of the placebo period. The small German open-label study with a 3-month treatment interruption after 12 months (Heuckmann *et al.* 1999) similarly observed a deterioration during the treatment interval. Such findings seem to point at the need for continuous treatment.

### *Maintenance of improvement*

Some studies report in more detail on the course of lung function changes during treatment. In several cases, an initial improvement has been followed by a subsequent deterioration. De Vuyst and colleagues (2000) report an 18-month follow-up of 33 patients with moderate disease. An increase in FVC of 8.5% at 6 weeks was followed by a decrease during which baseline values were reached at about 24 weeks, after which FVC remained stable. Similarly, $FEV_1$

increased by 14.4% in 6 weeks and reached baseline values at the 60th week, remaining stable thereafter. A small 6-month study of 20 patients with mild to moderate CF reported that overall, both FVC and $FEV_1$ had improved at the 1-month assessment, but fallen again thereafter (FVC at baseline: 60.7±2.5% predicted, at 3 months 70.4±2.8%, and at 6 months 63.3±4.2%) (Blau *et al.* 1997). Hence, such reports suggest that the extent of an initial favourable response may be reduced during ongoing therapy. However, other studies have reported continuous rises of the relevant lung function parameters over 12 months (Heuckmann *et al.* 1999).

### *Side effects*

The largest trial (Fuchs *et al.* 1994) reports that the administration of dornase alfa was associated with voice alterations (mainly hoarseness), pharyngitis and laryngitis, but no anaphylaxis. These types of side effects have been reported in many other studies. In an open-label study of similar size (974 patients), the most common treatment-related adverse events, voice alterations and pharyngitis, were each experienced by around 16% of the sample (Harms *et al.* 1998a).

### *Compliance*

The administration of dornase alfa via a nebuliser requires a significant time commitment from patients. Whereas this may be convenient in hospital, it may be less so as part of a busy lifestyle at home or work. A Czech study based on anonymous questionnaires found that 52.3% (45 patients) of responding CF patients reported that they omitted their inhalations occasionally; a further 8.1% (seven patients) forgot more often (Bartošová *et al.* 2000). A small study in Manchester assessed "compliance" by comparing the expected cost of dornase alfa treatment for 10 children and 13 adults with the actual recorded cost of encashed GP prescriptions for the drug (Talbot *et al.* 1998). Of course, encashed prescriptions do not equate directly to compliance, but they can be assumed to come closer to the truth than mere

prescription records themselves. Of the adult and child patients, 59% and 78% of the expected cost respectively were actually incurred. Suri and colleagues (2001) had observed 84% compliance in a 3-month trial of dornase alfa involving 48 children.

A study by Ollendorf *et al.* (2000) used healthcare claims over 30 months of dornase alfa treatment to estimate costs of care associated with different levels of use of dornase alfa. Interestingly, the median number of therapy days over the period was only 355. The 12 patients below median use showed a median of 180 days, the other half of the patients used it for a median of 494 days, still well below 100%-use. This small study points at the possibility that low use may be associated with higher costs of respiratory care, whereas prolonged use may reduce such costs; an annual saving of $2,500 was contrasted with an increase in the low-use group of $17,000, compared to costs calculated on the basis of a 6-month period preceding the initiation of dornase alfa.

### *Treatment practices*

Early data from the ERCF (1994 and 1995) indicates that 25% of the then 6,858 registered patients were receiving dornase alfa on enrolment, with only 16% in the UK (3,433 UK patients were then registered) (Koch *et al.* 1997). However, the intensity of and indications for use seem to still vary considerably. Whereas some see the main reason for prescribing dornase alfa to be for immediate symptom benefit (Geddes & Shah 1999), others may envisage a more or less continuous daily use at least in some patients.

Given the varied response as well as the cost of treatment, many centres have developed treatment initiation protocols, at times in co-operation with their respective purchasing authorities. Bradley and colleagues (2001) surveyed 49 CF centres in the UK and Ireland. Of the 42 responding centres, 22 had a standard written protocol for the commencement of dornase alfa. Indications

most commonly used for the initiation of treatment are: age >5 years, thick purulent sputum, difficulty expectorating sputum, deteriorating lung function, and recurrent infections. Outcome measures to assess the treatment's effectiveness include respiratory function, sputum production, pulse oximetry, and subjective awareness over an agreed trial period. The duration of individual trial periods on dornase alfa varies between centres from 2 weeks to 3 months (Geddes & Shah 1999). There are indications from some studies, that long-term response may be predicted on the basis of response shown after 3 months (Davies *et al.* 1997; Wizla-Derambure *et al.* 1998).

Some authors advocate formal "n-of-1 trials", incorporating randomisation, double-blinding and placebo control periods, to determine the individual response of any patient eligible for treatment (Innes 1998). Clinicians from Leeds for example reported their use of a 2-week trial period (longer for severe patients), after which the efficacy of dornase alfa is assessed by a minimum 10%-increase in $FEV_1$ or FVC from baseline and by subjective impression (Conway 1997). The adult centre at Liverpool uses an improvement in $FEV_1$ of ≥10% as the criterion for response (Ledson *et al.* 1998).

## Summary

- There are indications that benefits of dornase alfa in CF in terms of a reduced lung function decline and a reduced risk of respiratory exacerbations may continue over the long term. Randomised controlled trial evidence of the efficacy of dornase alfa is now available for a follow-up period of 2 years. In the longest (two-year) RCT, the treatment group exhibited a significantly reduced lung function decline vis-à-vis the control group. However, after one year, this difference was not significant (Quan *et al.* 2001).
- Longer-term RCTs do not exist and are very unlikely to be possible in future. Observational studies of more than two years follow-up are rare; they are so far of relatively small scale and frequently without comparison groups. The authors of a relatively recent four-year case-control study recommended further long-term studies involving larger cohorts (Shah *et al.* 2001).
- Considerable variations in the nature and extent of response between and probably also within individuals as well as practice variations between different CF centres add to the difficulties in evaluating the effectiveness of dornase alfa.

# *Data Entry Forms Used by the ERCF*

# Epidemiologic Registry of Cystic Fibrosis
## FORM 1 A - PATIENT ENROLLMENT

**Date of Enrollment:** DD MM YY **Centre #:** **Patient #:**

**Physician:**

## DEMOGRAPHICS

**Sex:** ☐ Male ☐ Female **Date of birth:** DD MM YY

Race/Ethnicity (*check all that apply*):

☐ Caucasian ☐ African

☐ Other (*specify*): _____

## CF DIAGNOSIS

**ear of CF diagnosis:** 1 9 ☐ ☐

iagnosis suggested by (*check all that apply*):

] Clinical symptoms: ☐ Meconium ileus ☐ Malnutrition/-absorption ☐ Pulmonary infection

Screening

Family history

sitive sweat test: ☐ Yes ☐ No ☐ Unknown or never tested

s patient been genotyped?

☐ Yes → *If yes, check mutations:*

☐ No

| | Chromosome #1 | Chromosome #2 |
|---|---|---|
| ΔF508 | ☐ | ☐ |
| Other (*specify*): _____ | ☐ | ☐ |
| Other (*specify*): _____ | ☐ | ☐ |
| Unknown: | ☐ | ☐ |

## ROUTINE THERAPIES ☐ check if none

ck all regularly scheduled medications or therapies
r than antibiotics for respiratory conditions) currently
ving or prescribed at time of enrollment:

- Airway clearance
- echnique/CPT
- Regular exercise
- Bronchodilator (oral)
- Bronchodilator (inhaled)
- Contraceptive (oral/implant)
- orticosteroid (oral)
- orticosteroid (inhaled)
- ther antiallergic therapy
- uretic
- sulin
- al hypoglycaemic

☐ Mucolytic/Expectorant
☐ NSAID

**Nutrition:**
☐ Oral supplements
☐ Tube feeding
☐ Parenteral

☐ Oxygen
☐ Assisted ventilation
☐ Pancreatic enzymes
☐ Other (*specify*): _____

## CLINICAL STATUS

Enrollment at time of: ☐ Clinic visit ☐ Hospitalisation (complete Form 4)

**If clinic visit, purpose of visit** (*check one*):
☐ Routine ☐ Exacerbation ☐ Research study
☐ New patient

**Height** ☐☐☐ cm **Weight** ☐☐☐.☐ kg

Cough frequency at time of enrollment (*check one*):
☐ None ☐ Occasionally ☐ Daily

Sputum productivity at time of enrollment (*check one*):
☐ None ☐ Occasionally ☐ Daily

If daily, estimate volume expectorated:
☐ <1 tablespoon/ minimal ☐ 1 tablespoon - 2 cups/moderate ☐ > 2 cups/ a lot

Sputum colour: ☐ Clear ☐ Yellow ☐ Green

Physical findings at time of enrollment (*check all that apply*):
☐ Crepitations ☐ Wheezing ☐ Clubbing ☐ Hyperinflatio

Oxygen saturation: ☐☐☐ % ☐ Room air ☐ $O_2$ ☐ check oxim not d

$pCO_2$: ☐☐ mm Hg (Torr)

## MEDICAL HISTORY ☐ check if no

**Check all medical conditions/illnesses (other than exacerbations of RTIs) within the last 6 months:**

**Pulmonary:**
☐ Bronchial hyperreactivity
☐ Allergic bronchopulmonary aspergillosis
☐ Mycobacterial disease (treated)
☐ Pneumothorax
Haemoptysis:
☐ Scant (streaking)
☐ Submassive (<1 cup in 24h)
☐ Massive (≥1 cup in 24h)

**Liver/Gastrointestinal:**
☐ Distal intestinal obstruction syndrome
☐ Gastro-oesophageal reflux
☐ Gall bladder disease
☐ Portal hypertension
☐ Elevated LFTs (bilirubin, enzymes)

**ENT:**
☐ Nasal polyposis
☐ Sinusitis (symptomatic)

**Other:**
☐ Congestive heart failure
☐ Other (*specify*): _____

Has patient received an organ transplant? ☐ Yes ☐ No
If yes, in year 1 9 ☐ ☐
Organ transplant (*specify*): _____

# Epidemiologic Registry of Cystic Fibrosis
## FORM 2 - CLINIC VISIT

**Visit Date:** ☐☐ DD ☐☐ MM ☐☐ YY  **Centre #:** ☐☐☐☐  **Patient #:** ☐☐☐☐ - ☐☐☐

**Physician:** _____

## CLINICAL STATUS

Purpose of visit *(check one)*:
☐ Routine ☐ Exacerbation ☐ Research study

**Height** ☐☐☐ cm (if patient <18 years)  **Weight** ☐☐☐.☐ kg

Cough frequency since last visit *(check one)*:
☐ None ☐ Occasionally ☐ Daily

Sputum productivity since last visit *(check one)*:
☐ None ☐ Occasionally ☐ Daily

If daily, estimate volume expectorated:
☐ <1 tablespoon/ minimal ☐ 1 tablespoon - 2 cups/moderate ☐ > 2 cups/ a lot

Sputum colour: ☐ Clear ☐ Yellow ☐ Green

Physical findings (*check all that apply*):
☐ Crepitations ☐ Wheezing ☐ Clubbing ☐ Hyperinflation

Oxygen saturation: ☐☐☐ % ☐ Room air ☐ $O_2$ ☐ check if oximetry not done

$pCO_2$: ☐☐ mm Hg (Torr)

## MEDICAL CONDITIONS ☐ check if none

Check all medical conditions/illnesses (other than exacerbations of RTIs) having emerged since last visit:

**Pulmonary:**
☐ Bronchial hyperreactivity
☐ Allergic bronchopulmonary aspergillosis
☐ Mycobacterial disease (treated)
☐ Pneumothorax
Haemoptysis:
☐ Scant (streaking)
☐ Submassive (<1 cup in 24h)
☐ Massive (≥1 cup in 24h)

**Liver/Gastrointestinal:**
☐ Distal intestinal obstruction syndrome
☐ Gastro-oesophageal reflux
☐ Gall bladder disease
☐ Portal hypertension
☐ Elevated LFTs (bilirubin, enzymes)

**ENT:**
☐ Nasal polyposis
☐ Sinusitis (symptomatic)

**Other:**
☐ Congestive heart failure
☐ Organ transplant (*specify*): _____
☐ Other (*specify*): _____

Coordinator: _____ Date: _____

Roche/Genentech Associate: _____ Date: _____

## DORNASE ALFA THERAPY

Indicate status of dornase alfa therapy since last visit:
☐ Not prescribed
☐ Prescription continued
☐ Prescribed at this visit
☐ Prescribed between visits → Start Date: ☐☐ DD ☐☐ MM ☐
☐ Prescription discontinued → Stop Date: ☐☐ DD ☐☐ MM ☐
Reason for discontinuation: _____

Current dornase alfa regimen:
Dose: ☐☐.☐ mg (1 ampoule dornase alfa = 2.5 m
Number of doses per day: ☐
If one dose per day: ☐ a.m. ☐ p.m.
Usage: ☐ Daily ☐ Other(*specify*): _____
Dornase alfa in relation to physiotherapy (PT): ☐ Dornase alfa before PT ☐ Dornase alf after PT
Type of nebulizer/compressor: _____ / _____
Indicate the number of prescribed doses missed in the last 7 days: ☐☐

## ROUTINE THERAPIES ☐ check if n

Check all regularly scheduled medications or therapies (other than antibiotics for respiratory conditions) currently receiving or have been prescribed at time of this clinic visit:

☐ Airway clearance technique/CPT
☐ Regular exercise
☐ Bronchodilator (oral)
☐ Bronchodilator (inhaled)
☐ Contraceptive (oral/implant)
☐ Corticosteroid (oral)
☐ Corticosteroid (inhaled)
☐ Other antiallergic therapy
☐ Diuretic
☐ Insulin
☐ Oral hypoglycaemic
☐ Mucolytic/Expectora
☐ NSAID

**Nutrition:**
☐ Oral supplements
☐ Tube feeding
☐ Parenteral

☐ Oxygen
☐ Assisted ventilation
☐ Pancreatic enzymes
☐ Other (*specify*): _____

## ADDITIONAL FORMS

Has a spirometry or full PFT been performed since last visit?
☐ Yes → ***Complete Form 3, Section A*** (best two results since last visit)
☐ No

Has patient had a respiratory culture obtained since last visit?
☐ Yes → ***Complete Form 3, Section B*** (most recent culture or obtained at current v
☐ No

Has patient received i.v., inhaled, or oral antibiotics for respiratory conditions since last visit?
☐ Yes → ***Complete Form 3, Section C***
☐ No

Has patient had white blood cell count and serum IgG determined since last visit?
☐ Yes → ***Complete Form 3, Section D*** (most recent values or obtained at current v
☐ No

Has patient experienced a serious adverse event or had a non-routine hospitalisation since last visit?
☐ Yes → ***Complete Form 4***
☐ No

# Epidemiologic Registry of Cystic Fibrosis

## FORM 3 - PULMONARY FUNCTION TESTS, MICROBIOLOGY, ANTIBIOTICS, WBC, IgG

**Visit Date:** DD MM YY **Centre #:** **Patient #:** -

**Physician:**

### A. PULMONARY FUNCTION TESTS (Actual)

☐ check if no

| Date of Test DD | MM | YY | Height (cm) (if <18 years) | FVC (l) | $FEV_1$ (l) | $FEF_{25-75\%}$ (l/sec) | RV (l) | TLC (l) |
|---|---|---|---|---|---|---|---|---|
| | | | | . | . | . | . | . |
| | | | | . | . | . | . | . |

### B. MICROBIOLOGY

☐ check if no

| Sampling Date DD | MM | YY | Type of Culture | Organisms Present: Codes* (circle all that apply) | Other |
|---|---|---|---|---|---|
| | | | ☐ Sputum ☐ Throat ☐ BAL | 0 1 2 3 4 5 6 7 8 9 10 11 12 ⟶ *specify:* | |

* **Note:**
0. None/Normal flora
1. P. aeruginosa (non-mucoid)
2. P. aeruginosa (mucoid)
3. P. aeruginosa (multiply resistant)
4. P. cepacia
5. Other Pseudomonas
6. Staph. aureus
7. H. influenzae
8. Xanthomonas
9. Candida
10. Aspergillus
11. M. tuberculosis
12. Other (*specify on above line*)

### C. ANTIBIOTICS

☐ check if no

Complete section if patient is currently (including prescription of this clinic visit) receiving or has been prescribed since last visit: **i.v., inhaled, or oral antibiotics** for respiratory conditions.

**CODING INSTRUCTIONS**

**Indication:** 1 = Prophylaxis, continuous use; 2 = Prophylaxis, intermittent use; 3 = Exacerbation of RTI

**Route:** 1 = I.v. in hospital; 2 = I.v. at home; 3 = Inhaled; 4 = Oral

| | Antibiotic Name (Generic Name) | Indication Code | Route Code | Start Date — Check if start date unknown ✓ | DD | MM | YY | Stop Date — Check if continuing ✓ | DD | MM | YY |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | | | | ☐ | | | | ☐ | | | |
| 2. | | | | ☐ | | | | ☐ | | | |
| 3. | | | | ☐ | | | | ☐ | | | |
| 4. | | | | ☐ | | | | ☐ | | | |
| 5. | | | | ☐ | | | | ☐ | | | |
| 6. | | | | ☐ | | | | ☐ | | | |

### D. WHITE BLOOD CELL COUNT AND SERUM IgG

☐ check if no

| Date of Test DD | MM | YY | WBC ($10^9$/l) | Date of Test DD | MM | YY | IgG (g/l) |
|---|---|---|---|---|---|---|---|
| | | | . | | | | . |

Coordinator: _____ Date: _____

Roche/ Genentech Associate: _____ Date: _____

# Epidemiologic Registry of Cystic Fibrosis
## FORM 2 - CLINIC VISIT

**Visit Date:** ☐☐ DD ☐☐ MM ☐☐ YY **Centre #:** ☐☐☐☐ **Patient #:** ☐☐☐☐ - ☐☐☐

**Physician:** _____

## CLINICAL STATUS

Purpose of visit *(check one)*:
☐ Routine ☐ Exacerbation ☐ Research study

**Height** ☐☐☐ cm **Weight** ☐☐☐.☐ kg
(if patient <18 years)

Cough frequency since last visit *(check one)*:
☐ None ☐ Occasionally ☐ Daily

Sputum productivity since last visit *(check one)*:
☐ None ☐ Occasionally ☐ Daily

If daily, estimate volume expectorated:
☐ <1 tablespoon/ minimal ☐ 1 tablespoon - 2 cups/moderate ☐ > 2 cups/ a lot

Sputum colour: ☐ Clear ☐ Yellow ☐ Green

Physical findings (*check all that apply*):
☐ Crepitations ☐ Wheezing ☐ Clubbing ☐ Hyperinflation

Oxygen saturation: ☐☐☐ % ☐ Room air ☐ $O_2$ ☐ check if oximetry not done

$pCO_2$: ☐☐ mm Hg (Torr)

## MEDICAL CONDITIONS ☐ check if none

Check all medical conditions/illnesses (other than exacerbations of RTIs) having emerged since last visit:

**Pulmonary:**
☐ Bronchial hyperreactivity
☐ Allergic bronchopulmonary aspergillosis
☐ Mycobacterial disease (treated)
☐ Pneumothorax
Haemoptysis:
☐ Scant (streaking)
☐ Submassive (<1 cup in 24h)
☐ Massive (≥1 cup in 24h)

**Liver/Gastrointestinal:**
☐ Distal intestinal obstruction syndrome
☐ Gastro-oesophageal reflux
☐ Gall bladder disease
☐ Portal hypertension
☐ Elevated LFTs (bilirubin, enzymes)

**ENT:**
☐ Nasal polyposis
☐ Sinusitis (symptomatic)

**Other:**
☐ Congestive heart failure
☐ Organ transplant (*specify*): _____
☐ Other (*specify*): _____

Coordinator: _____ Date: _____

Roche/Genentech Associate: _____ Date: _____

## DORNASE ALFA THERAPY

Indicate status of dornase alfa therapy since last visit:
☐ Not prescribed
☐ Prescription continued
☐ Prescribed at this visit
☐ Prescribed between visits → Start Date: ☐☐ DD ☐☐ MM ☐
☐ Prescription discontinued → Stop Date: ☐☐ DD ☐☐ MM ☐
Reason for discontinuation: _____

Current dornase alfa regimen:
Dose: ☐☐.☐ mg (1 ampoule dornase alfa = 2.5 m
Number of doses per day: ☐
If one dose per day: ☐ a.m. ☐ p.m.
Usage: ☐ Daily ☐ Other(*specify*): _____
Dornase alfa in relation to physiotherapy (PT): ☐ Dornase alfa before PT ☐ Dornase alf after PT
Type of nebulizer/compressor: _____ / _____
Indicate the number of prescribed doses missed in the last 7 days: ☐☐

## ROUTINE THERAPIES ☐ check if r

Check all regularly scheduled medications or therapies (other than antibiotics for respiratory conditions) currently receiving or have been prescribed at time of this clinic visit:

☐ Airway clearance technique/CPT
☐ Regular exercise
☐ Bronchodilator (oral)
☐ Bronchodilator (inhaled)
☐ Contraceptive (oral/implant)
☐ Corticosteroid (oral)
☐ Corticosteroid (inhaled)
☐ Other antiallergic therapy
☐ Diuretic
☐ Insulin
☐ Oral hypoglycaemic
☐ Mucolytic/Expectora
☐ NSAID

**Nutrition:**
☐ Oral supplements
☐ Tube feeding
☐ Parenteral

☐ Oxygen
☐ Assisted ventilation
☐ Pancreatic enzymes
☐ Other (*specify*): _____

## ADDITIONAL FORMS

Has a spirometry or full PFT been performed since last visit?
☐ Yes → ***Complete Form 3, Section A*** (best two results since last visit)
☐ No

Has patient had a respiratory culture obtained since last visit?
☐ Yes → ***Complete Form 3, Section B*** (most recent culture or obtained at current v
☐ No

Has patient received i.v., inhaled, or oral antibiotics for respiratory conditions since last visit?
☐ Yes → ***Complete Form 3, Section C***
☐ No

Has patient had white blood cell count and serum IgG determined since last visit?
☐ Yes → ***Complete Form 3, Section D*** (most recent values or obtained at current v
☐ No

Has patient experienced a serious adverse event or had a non-routine hospitalisation since last visit?
☐ Yes → ***Complete Form 4***
☐ No

# Epidemiologic Registry of Cystic Fibrosis
## FORM 4 - HOSPITALISATION / SERIOUS ADVERSE EVENT

**Date of Report:** ☐☐ DD ☐☐ MM ☐☐ YY  **Centre #:** ☐☐☐☐  **Patient #:** ☐☐☐☐ - ☐☐☐☐

**Physician:** _____

**Sex:** ☐ Male ☐ Female  **Birth date:** ☐☐ DD ☐☐ MM ☐☐ YY

## SERIOUS ADVERSE EVENT INCLUDING NON-ROUTINE HOSPITALISATION

**Most significant event/illness or reason for hospitalisation:**

Onset of event: ☐☐ DD ☐☐ MM ☐☐ YY

Event resolved? ☐ Yes ☐ No  Date of resolution: ☐☐ DD ☐☐ MM ☐☐ YY

**Describe event** (including relevant medical history, tests, and laboratory data).

**Outcomes attributed to serious adverse event** (*check all that apply*):

☐ Hospitalisation

Date of admission: ☐☐ DD ☐☐ MM ☐☐ YY

Date of discharge: ☐☐ DD ☐☐ MM ☐☐ YY

☐ Prolonged hospitalisation

☐ Life-threatening

☐ Required intervention (to prevent permanent impairment or damage)

☐ Permanent disability

☐ Other (*specify*): _____

☐ Death → Date: ☐☐ DD ☐☐ MM ☐☐ YY (*complete Form 5*)

Cause(s) of death:

☐ Cardio-respiratory failure

☐ Other (*specify*): _____

☐ Other (*specify*): _____

☐ Other (*specify*): _____

**Cause of event** (*check all that apply*):

☐ Complications of CF

☐ Dornase alfa* (*See note below*)

☐ Other medication (*specify*): _____

☐ Other (*specify*): _____

☐ Unknown

**Was patient receiving dornase alfa therapy at time of event?** ☐ Yes ☐ No

If yes, is event related to dornase alfa*? ☐ Yes ☐ No

*If yes, please complete:*

Dornase alfa initiation date: ☐☐ DD ☐☐ MM ☐☐ YY

Dose: ☐☐.☐ mg (1 ampoule dornase alfa = 2.5 mg)

Number of doses per day: ☐

**Change in dornase alfa therapy** (*if applicable, check one*):

☐ No change
☐ Reduced dose
☐ Increased dose
☐ Interrupted
☐ Discontinued

*** To physician: If you consider this event related to dornase alfa therapy, sign form below and immediately FAX to Quintiles +49-6102-296 296**

Physician's signature: _____  FAX date: ☐☐ DD ☐☐ MM ☐☐ YY

Coordinator: _____  Date: _____

Telephone Number: _____ / _____ / _____ (Country / Area / Extension)

Roche/Genentech Associate: _____  Date: _____

# Epidemiologic Registry of Cystic Fibrosis

## NON-SERIOUS ADVERSE DRUG REACTION REPORT FORM

**Date of Report:** DD MM YY **Centre #:** **Patient #:**

**Physician:**

### DEMOGRAPHY

**Sex:** ☐ Male ☐ Female **Date of birth:** DD MM YY **Height** cm **Weight** kg

### DORNASE ALFA (PULMOZYME) DOSAGE

**Start date:** DD MM YY

Dose: . mg (1 ampoule dornase alfa = 2.5 mg) Number of doses per day:

Subsequent dosage adjustment ( *specify*):

### ADVERSE DRUG REACTION

Onset of reaction: DD MM YY Reaction resolved? ☐ Yes ☐ No Date of resolution: DD MM YY

Describe reaction (including relevant medical history, tests, and laboratory data).

### CHANGE IN DORNASE ALFA (PULMOZYME) THERAPY

**Change in dornase alfa (Pulmozyme) therapy** (*check one*): ☐ No change ☐ Increased dose ☐ Discontinued ☐ Reduced dose ☐ Interrupted

Date of dose change: DD MM YY

### CONCOMITANT MEDICATION

Please list with dosages, dates, routes and indications:

### CAUSAL RELATIONSHIP

Causal relationship (drug reaction): ☐ Remote ☐ Possible ☐ Probable

Have you reported this ADR to the CSM/NDAB? ☐ Yes ☐ No Date of report: DD MM YY

**To physician: Please return the top copy of this completed form to the Drug Surveillance Department, Roche Products Ltd., in the reply-paid envelope provided.**

Physician's signature: Date:

# *Response letter from Data Protection Commissioner*

# DATA PROTECTION

Ms Judith Strobl,
Research Fellow,
Prescribing Research Group,
Dept. of Pharmacology and Therapeutics,
The Infirmary,
The University of Liverpool,
70 Pembroke Place,
Liverpool L69 3GF.

Our Ref: G0312/IM
17th March 2000

Dear Ms Strobl,

Thank you for your letter of 16/2/00. As I understand it you have two questions:

1) Is patient consent needed to use coded or anonymised data for your research ?

2) If "internal" staff from within the participating CF units/NHS Trusts which supplied the anonymised data to the ERCF database "de-anonymise" the data in order to validate it within the Trust and given that only those staff will have access to any data in a patient identifiable format is it necessary to obtain individual patient consent to this use of the data ?

The Data Protection Act 1998 defines personal data as, "...data which relate to a living individual who can be identified from those data, or from those data and other information which is in the possession of, or is likely to come into the possession of the data controller....".

The First Principle of the Data Protection Act 1998 requires, amongst other things, that a data controller, e.g a NHS Trust, shall process personal data fairly. Fairness includes ensuring that data subjects i.e. patients in this case, have an understanding of the purposes for which their personal data are to be processed, the likely consequences of such processing and, more particularly, whether particular disclosures of their personal data can be envisaged.

It is the Data Protection Commissioner's view that even "anonymised" personal data will remain personal data to those who only have the encoded information extracted from it, as well as to whoever holds the encoding key, provided the key remains in existence. The fact that at some future point apparently "anonymised" data held by a particular person could be "re-identified" means that the data will still constitute personal data whilst held in an "anonymised" form by that person. In such a scenario we take the view that there has not been true anonymisation but rather something more akin to "pseudonymisation" of what remains personal data subject to the Act.

It appears from the information you have provided that the encoded data on the ERCF database is still personal data as defined by the Act because the key is still in the possession of the CF units/NHS Trusts which supplied the personal data.

# DATA PROTECTION

Having said that I should stress that from the point of view of the Data Protection Commissioner this does not mean that such personal data cannot be processed in compliance with the Act. Whilst it may be very difficult to ever truly anonymise data it is comparatively easy to ensure that what is, in effect, pseudonymised personal data, is processed fairly and lawfully.

To do this you will need to ensure that you have a well constructed security protocol in place between yourselves and the CF units/NHS Trusts which ensures the pseudonymity of the personal data and prevents anyone, other than the appropriate staff within the CF units/NHS Trusts, from having access to the encoding key.

In addition, in order to comply with the First Principle of the Data Protection Act 1998, the Trusts should have arrangements in place to inform patients of the likely uses of their personal data, including the fact that it may be disclosed for research purposes, and to deal with any objections from patients. The provision of information to patients in this way is not something that the Data Protection Act 1998 has introduced for the first time but is something that the Trusts should already have been doing in order to comply with the 1984 Data Protection Act.

However, the proper provision of patient information has not always been treated as a priority in some NHS units. The Data Protection Act 1998 brings this issue into sharper focus, placing as it does, more stringent legal requirements on data controllers to provide specific information to data subjects and to ensure that they are able to signify their agreement to their personal data being processed. If any of the six NHS Trusts participating in your research do not currently have adequate arrangements in place for the proper provision of patient information then they should introduce these as soon as possible.

We recognise the practical difficulties of doing this for patients whose personal data is already on the database and would not expect the Trusts to try and obtain consent in retrospect. However in the event of a request from a data subject whose personal data is already on the database, the Data Protection Commissioner reserves the right to assess whether or not their personal data has beeen processed in compliance with the Act.

I hope you find my response of value, if you wish to discuss the matter in more detail you can reach me on 01625 545712.

Yours sincerely,

Ian Miller,
Compliance Officer.

# *Formulae for predicted lung function used in the study*

Conventions: " * " = multiplication ; " ^ " = (raise to) power

**"Dundee" (UK CFF) formulae**
per website www.cystic-fibrosis.org.uk

For $FEV_1$:

Age under 18 years (height in m)
("From Polgar, Pulmonary Function Testing in Children ")
Male: 0.812 * Height ^ 2.77
Female: 0.788 * Height ^ 2.73

18 and above (height in cm)
("From Reuben M Cherniak, Pulmonary Function Testing ")
Male: 0.04525 * Height - 0.03509 * Age - 2.59946
Female: 0.04071 * Height - 0.02147 * Age - 2.56958.

For FVC:

Age less than 18 years (height in m)
("From Polgar, Pulmonary Function Testing in Children ")
Male: 1.004 * Height ^ 2.72
Female: 0.946 * Height ^ 2.61

18 and above (height in cm):
("From Reuben M Cherniak, Pulmonary Function Testing ")
Male: (0.06584 * Height - 0.02954) * Age) – 5.12451
Female: (0.05557 * Height) – (0.00793 * Age) – 4.89036.

**"Knudson" formulae**

Source: Knudson et al (1983) Changes in the normal expiratory flow volume curve with growth and ageing, *American Review of Respiratory Disease*, **127** 725 - 734

For $FEV_1$:
Heights in centimetres

| | |
|---|---|
| Males | |
| Ages 6 but less than 12 | (0.0348*height) – 2.8142 |
| Age 12 but less than 25 | (0.0519*height) + (age*0.0636) – 6.1181 |
| Age 25 and over | (0.0665*height) + (age*0.0292) – 6.5147 |
| Females | |
| Ages 6 but less than 11 | (0.0336*height) – 2.7578 |
| Age 11 but less than 20 | (0.0351*height) + (age*0.0694) – 3.7622 |
| Age 20 and over | (0.0309*height) + (age*0.0201) – 1.4050 |

For FVC:
Heights in centimetres

| | |
|---|---|
| Males | |
| Ages 6 but less than 12 | (0.0409*height) – 3.3756 |
| Age 12 but less than 25 | (0.0590*height) + (age*0.0739) – 6.8865 |
| Age 25 and over | (0.0844*height) + (age*0.0298) – 8.7818 |
| Females | |
| Ages 6 but less than 11 | (0.0430*height) – 3.7486 |
| Age 11 but less than 20 | (0.0416*height) + (age*0.0699) – 4.4470 |
| Age 20 and over | (0.0427*height) + (age*0.0174) – 2.9001 |

## *Sample forms for sample-DQR*

# Cost-Effectiveness of rhDNase Study Group

**Data Quality Review**

### Section II/1: Demographic Data

| | |
|---|---|
| **Patient ID:** | xxx |
| **Current treatment centre:** | xxx |

### Section 1: Demographic Data

| Data item | Current entry | Corrections and Comments |
|---|---|---|
| Sex | xxx | |
| Date of birth (month and year) | xxx | |
| Year of CF diagnosis | xxx | |
| Has patient been genotyped? | xxx | |
| Chromosome 1 mutation | xxx | |
| Chromosome 2 mutation | xxx | |
| Enrolment date | xxx | |
| Discontinuation date | xxx | |
| Reason for discontinuation | xxx | |

Comments:

# Cost-Effectiveness of rhDNase Study Group: Data Quality Review
# Section II/2: Lung function

| | |
|---|---|
| **Patient ID:** | **xxx** |
| **Current treatment centre:** | **xxx** |

***Please check details of lung function recorded at each reporting date and NOTE ANY ERRORS in right-hand column (all measures are in actual litres)***

| **Date of report** | Test date | Type of record | If enrolment record, is result from time PRIOR to rhDNase use? | **FVC** (Forced Vital Capacity) (litres) | **FEV1** (Forced Expiratory Volume in 1 second) (litres) | **Errors/Corrections PLEASE TICK if data in row is correct** |
|---|---|---|---|---|---|---|
| *19-Oct-94* | *19-Oct-94* | *enrolment* | *no* | *2.08* | *1.53* | |
| *08-Sep-95* | *8-Sep-95* | *follow-up* | | *1.92* | *1.30* | |
| *17-Nov-95* | *17-Nov-95* | *follow-up* | | *1.99* | *1.44* | |
| *19-Jan-96* | *19-Jan-96* | *follow-up* | | *2.41* | *1.77* | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

# Summary List 1 - current or most recent patients

## *centre name*

| bpid | Enrolment date | Centre (at enrolment) | Transfer date | New centre | Current centre | Last reporting date | Date of death | Cause of death | Discontinuation date | CORRECTIONS |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |

## Summary List 2 - Patients enrolled at centre, but transferred away

***centre name***

| | bpid | Enrolment date | Centre (at enrolment) | Transfer date | New centre | Current centre | CORRECTIONS |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

# *Further Baseline Data*

The following baseline data presentation is limited to the 999 patients in the study sample (see Figure 4.2 for selection of sample).

## *Age, sex*

The sample is reasonably well balanced in terms of sex and age groupings, with a slightly larger proportion of males across all age groups, particularly the adult group (Table A.3).

**Table A.3: Sex and age at enrolment**

| Age at enrolment | Male | | Female | | Total | |
|---|---|---|---|---|---|---|
| <6 years | 88 | 8.8% | 65 | 6.5% | 153 | **15.3%** |
| 6 to <13 years | 136 | 13.6% | 129 | 12.9% | 265 | **26.6%** |
| 13 to <18 years | 111 | 11.1% | 98 | 9.8% | 209 | **20.9%** |
| 18 years and over | 206 | 20.6% | 165 | 16.5% | 371 | **37.2%** |
| **Total** | **541** | **54.2%** | **457** | **45.8%** | **998** | **100.0%** |

Enrolment date of one male patient unknown.

## *Genotype*

Only 729 patients could be clearly categorised according to genotype. Centres 1 and 5 - children's and adult centres in the same conurbation - showed a markedly higher proportion of patients with less severe genotypes (Table A.4). Patients with less severe genotypes showed higher lung function on enrolment (Table A.5) and were much less likely to use dornase alfa (Table A.6).

**Table A.4: Severity of genotype by enrolment centre**

| Enrolment centre | A (severe) | | B (less severe) | | Total |
|---|---|---|---|---|---|
| 1 | 104 | 93% | 8 | 7% | 112 |
| 2 | 62 | 97% | 2 | 3% | 64 |
| 3 | 128 | 97% | 4 | 3% | 132 |
| 4 | 108 | 98% | 2 | 2% | 110 |
| 5 | 59 | 89% | 7 | 11% | 66 |
| 6 | 154 | 96% | 7 | 4% | 161 |
| 7 | 82 | 98% | 2 | 2% | 84 |
| Total | 697 | 96% | 32 | 4% | 729 |

**Table A.5: Severity of genotype by sex and lung function on enrolment**

| Sex | $FEV_1$ (% pred.) | A (severe) | | B (less severe) | | Total |
|---|---|---|---|---|---|---|
| Male | < 40 | 70 | 97% | 2 | 3% | 72 |
| | 40 - 69.9 | 105 | 97% | 3 | 3% | 108 |
| | 70+ | 150 | 94% | 10 | 6% | 160 |
| All male | | 325 | 96% | 15 | 4% | 340 |
| Female | < 40 | 44 | 96% | 2 | 4% | 46 |
| | 40 - 69.9 | 100 | 96% | 4 | 4% | 104 |
| | 70+ | 118 | 94% | 8 | 6% | 126 |
| All female | | 262 | 95% | 14 | 5% | 276 |
| **Total** | | 587 | 95% | 29 | 5% | 616 |

**Table A.6: Dornase alfa use by severity of genotype**

| Any dornase alfa use recorded | Genotype category | | | |
|---|---|---|---|---|
| | A (severe) | | B (less severe) | |
| No | 373 | 53.5% | 27 | 84.4% |
| Yes | 324 | 46.5% | 5 | 15.6% |
| Total | 697 | 100.0% | 32 | 100.0% |

### *Diabetes*

Diabetes was diagnosed or treatment of diabetes (insulin or oral anti-diabetic drugs) was recorded in a total of 147 (14.7%) patients. Diabetics in the sample were more likely to be female (Table A.7), have poor lung function on enrolment (Table A.8), and have been receiving dornase alfa, compared to non-diabetics (Table A.9).

**Table A.7: Diabetes by sex**

| Sex | No record of diabetes | | Diabetes* | | Total |
|---|---|---|---|---|---|
| Male | 476 | 87.8% | 66 | 12.2% | 542 |
| Female | 376 | 82.3% | 81 | 17.7% | 457 |
| Total | 852 | 85.3% | 147 | 14.7% | 999 |

*Diagnosis or treatment of diabetes recorded.

**Table A.8: Diabetes by lung function on enrolment**

| $FEV_1$ (%pred) | No record of diabetes | | Diabetes* | | Total |
|---|---|---|---|---|---|
| < 40 | 119 | 68.0% | 56 | 32.0% | 175 |
| 40 - 69.9 | 242 | 84.3% | 45 | 15.7% | 287 |
| 70+ | 333 | 89.5% | 39 | 10.5% | 372 |
| Total | 694 | 83.2% | 140 | 16.8% | 834 |

*Diagnosis or treatment of diabetes recorded.

**Table A.9: Diabetes by dornase alfa use**

| Any dornase alfa use recorded | No record of diabetes | | Diabetes* | |
|---|---|---|---|---|
| No | 499 | 58.6% | 53 | 36.1% |
| Yes | 353 | 41.4% | 94 | 63.9% |
| Total | 852 | 100.0% | 147 | 100.0% |

*Diagnosis or treatment of diabetes recorded.

### *Microbiological colonisation on enrolment*

Enrolment reports of 904 patients contained culture reports. In some of these patients, the relevant culture may have been taken some weeks or months before the actual enrolment date; hence this is not to be treated as strict point prevalence.

There are differences in reported colonisation rates on enrolment for all recorded micro-organisms between individual centres (see Table A.10). For example, rates of colonisation with *B.cepacia* vary between 2.8% and 4.8% in children's centres and as much as 8.1% and 25.6% in adult centres. Rates for *Pseudomonas* or *Xanthomonas* vary between 41.8% and 75.2% and 73.8% and 82.5% respectively. These large variations may put into question the completeness of recording of these variables, and again demonstrate the importance of between-centre variations.

**Table A.10: Microbiological colonisation on enrolment by enrolment centre**

| Enrolment centre | Patients with some colonisation reported | | No colonisation reported | | Patients with microbiology report |
|---|---|---|---|---|---|
| 1 | 123 | 98.4% | 2 | 1.6% | 125 |
| 2 | 88 | 89.8% | 10 | 10.2% | 98 |
| 3 | 109 | 81.3% | 25 | 18.7% | 134 |
| 4 | 122 | 85.9% | 20 | 14.1% | 142 |
| 5 | 79 | 96.3% | 3 | 3.7% | 82 |
| 6 | 161 | 98.8% | 2 | 1.2% | 163 |
| 7 | 144 | 90.0% | 16 | 10.0% | 160 |
| Total | 826 | 91.4% | 78 | 8.6% | 904 |

### *Dornase alfa use*

A total of 447 patients (44.7%) were reported ever to have received dornase alfa. The majority of patients were apparently commenced on treatment during 1994 and 1995. Children's centres appear to use dornase alfa less frequently than adult centres, but this may be partly due to the fact that some patients in the sample were too young during part of the follow-up period (Table A.11).

Twenty seven per cent of dornase alfa users have received the drug for a period of less than one year in total, regardless of interruptions, 4.4% (19 patients) for 14 or fewer days. For 105 patients (24% of dornase users) the cumulative treatment duration was 4 years or more. Nearly all patients were on 2.5 mg daily for most of their treatment period.

For 68 patients (15.2% of dornase users), the first available spirometry results post-dated the first treatment initiation date. For the remainder of patients, the first spirometry either coincided with the first treatment initiation date, predated it, or the dates were missing.

**Table A.11: Dornase therapy by enrolment centre**

| Enrolment centre | Continuous use | | | | Intermittent use | | | | No use | | Total (100%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Start before enrolment | | Start on or after enrolment | | Start before enrolment | | Start on or after enrolment | | | | |
| 1 | | | 32 | 24.6% | 1 | .8% | 11 | 8.5% | 86 | 66.2% | 130 |
| 2 | 14 | 13.9% | 10 | 9.9% | 12 | 11.9% | 3 | 3.0% | 62 | 61.4% | 101 |
| 3 | 17 | 10.8% | 28 | 17.8% | 7 | 4.5% | 6 | 3.8% | 99 | 63.1% | 157 |
| 4 | 17 | 11.9% | 15 | 10.5% | 1 | .7% | 24 | 16.8% | 86 | 60.1% | 143 |
| 5 | 2 | 2.3% | 10 | 11.6% | 9 | 10.5% | 25 | 29.1% | 40 | 46.5% | 86 |
| 6 | 32 | 15.0% | 47 | 22.0% | 9 | 4.2% | 28 | 13.1% | 98 | 45.8% | 214 |
| 7 | 28 | 16.8% | 45 | 26.9% | 5 | 3.0% | 9 | 5.4% | 80 | 47.9% | 167 |
| Total | 110 | 11.0% | 187 | 18.7% | 44 | 4.4% | 106 | 10.6% | 551 | 55.2% | 998 |

Table A.12 presents the colonisation of patients on enrolment broken down by dornase alfa use group. Clearly, any dornase alfa user group showed higher rates of colonisation with Pseudomonas, particularly those treated before enrolment. This corresponds well with the notion that at least in some centres a colonisation with Pseudomonas was amongst the criteria for selection for treatment with dornase alfa. Similarly, colonisation with *B. cepacia* was far higher in the dornase alfa treated groups. This trend was not visible for all micro-organisms.

**Table A.12: Per cent of patients with colonisation reported on enrolment by dornase alfa use group, (n=904 with microbiological cultures reported on enrolment report)**

| Dornase alfa use category | N | Pseudomonas, Xanthomonas | B. cepacia | S. aureus | Haemophilus | Any colonisation |
|---|---|---|---|---|---|---|
| No use | 492 | 56.1% | 4.1% | 32.3% | 23.2% | 87.0% |
| Continuous use - start before enrolment | 89 | 88.8% | 13.5% | 39.3% | 16.9% | 97.8% |
| Continuous use - start on or after enrolment | 185 | 78.9% | 16.2% | 32.4% | 16.8% | 96.8% |
| Intermittent use - start before enrolment | 39 | 87.2% | 17.9% | 38.5% | 33.3% | 97.4% |
| Intermittent use - start on or after enrolment | 99 | 79.8% | 24.2% | 23.2% | 19.2% | 94.9% |
| Total | 904 | 67.9% | 10.3% | 32.3% | 21.2% | 91.4% |

The main reported reasons for discontinuation of dornase therapy were lack of effectiveness, but also an intended short treatment period, most likely during a hospital stay (Table A.13).

**Table A.13: Reported reasons for discontinuation of dornase alfa**

| Reason | Occurrences (number and %) | |
|---|---|---|
| Lack of effectiveness stated | 60 | 28% |
| Planned as short-term use (e.g. in hospital) | 52 | 24% |
| Compliance / patient choice | 43 | 20% |
| Any symptoms mentioned | 31 | 14% |
| External reason | 12 | 6% |
| Trial stated as reason for discontinuation | 7 | 3% |
| Death | 7 | 3% |
| Dose change | 4 | 2% |
| Other | 4 | 2% |
| Unknown reason | 3 | 1% |
| Total | 218 | 100% |

Note: Patients may appear more than once, $n$=149.

## *Further Outcome Data*

### Outcome: Exacerbations

Table A.14 presents a subset of the main study sample with at least one year's observations available (n=857), as computations from short periods of observations are unstable, and high numbers of exacerbations over a short period may indicate problems of acute rather then chronic disease stage.

After annual rates had been computed and ranked, the data were dichotomised (into two equal parts). Dichotomisation was carried out separately for males and females to avoid possible confounding by sex.

The dichotomised data showed no difference between children and adults, possibly indicating that exacerbations were not related to the progression of the disease. Although hardly surprising, the association of dichotomised exacerbation data with death is statistically significant ($p<0.001$) and the rates for the death cases about 50% higher than survivors. The numbers are too small to seek general conclusions on increases in rate over time in the death cases.

Table A.15 illustrates a six-fold categorisation of rates of exacerbation - for convenience; the bands indicate the number of calendar quarters per year a "typical" patient was affected. The numbers show a continuum. The "no recorded dornase alfa group" had a high proportion of patients in the "no exacerbation" category, but still sizable numbers of patients in the highest categories.

The data are difficult to interpret. It is worth pointing out that the frequency of exacerbations reported varied between centres, and under-reporting in some of them seemed highly likely. In addition, the definition of the variable was problematic to say the least, given that three different points in the database indicated exacerbations, and the timely overlap between episodes was very unclear. Based on advice from clinicians, one week was considered the minimum time to separate two distinct episodes. Observed variations may have arisen at least in part due to differential impact of this definitional rule, or reporting differences between centres. This demonstrates that in ESDs, inconsistent reporting of outcomes and poor definition of key outcome variables can jeopardise any meaningful analysis.

**Table A.14: Summary of individual annualised exacerbation rates by sex, initial age group and dornase alfa user group: also dichotomised by broad age group (<18/18+), and diabetes status and categorical division (n= 857 with at least one year of observations) (7,867 exacerbations).**

Rates computed as (n exacerbations) / (n years of observation to nearest quarter)

| | | Mean exacerbation rates | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Initial five-year age group** | | 5<10 | 10<15 | 15<20 | 20<25 | 25<30 | 30<35 | 35+ | all |
| | Male | 2.52 | 2.01 | 2.20 | 2.63 | 2.51 | 2.46 | 2.60 | 2.40 |
| **Sex** | Female | 2.60 | 2.53 | 3.01 | 2.72 | 2.85 | 2.60 | 2.76 | 2.72 |
| | All | 2.55 | 2.32 | 2.60 | 2.66 | 2.72 | 2.51 | 2.67 | 2.55 |
| | n | 262 | 150 | 181 | 138 | 65 | 41 | 20 | 857 |

| | | Use of dornase alfa | | | | | |
|---|---|---|---|---|---|---|---|
| | | Continuous | | Some | | None | |
| | | Starting before enrolment | Starting on or after enrolment | Starting before enrolment | Starting on or after enrolment | | all |
| **Group number** | | 1 | 2 | 3 | 4 | 5 | |
| | Male | 2.71 | 3.01 | 2.33 | 2.77 | 2.11 | 2.40 |
| **Sex** | Female | 3.32 | 2.92 | 3.42 | 3.42 | 2.26 | 2.72 |
| | All | 3.03 | 2.96 | 2.86 | 3.06 | 2.17 | 2.55 |
| | n | 72 | 186 | 37 | 103 | 459 | 857 |

| **Dichotomised data (including "0")** | | **Age** | | **Diabetes** | |
|---|---|---|---|---|---|
| | | <18 | 18+ | No | Yes |
| **Ranking by annual rate** | Low | 275 | 154 | 365 | 64 |
| | High | 260 | 168 | 358 | 70 |
| | All | 535 | 322 | 695 | 134 |
| T-test for significance of imbalance in dichotomised data | | Male: p=0.32<br>Female: p=0.21 | | p=0.57, and no difference controlled for age and sex | |

| **Exacerbations in relation to deaths - dichotomised data** | | | | | |
|---|---|---|---|---|---|
| | | Survivors | Deaths | All | |
| **Ranking by annual rate** | Low | 417 | 12 | 429 | p<0.001 for male, female, and overall. |
| | High | 364 | 64 | 428 | |
| | All | 781 | 76 | 857 | |

| **Exacerbations in relation to deaths** | | | | |
|---|---|---|---|---|
| | Survivors | Deaths | All | |
| **Mean exacerbation rates** | Male | 2.28 | 3.70 | 2.40 |
| | Female | 2.53 | 4.50 | 2.72 |
| | All | 2.40 | 4.10 | 2.55 |

**Table A.15: Categorical division of annual rates of exacerbations and dornase alfa use group classification (n= 857 with at least one year of observations) (7,867 exacerbations).**

| | Use of dornase alfa | | | | | Totals | | |
|---|---|---|---|---|---|---|---|---|
| | Continuous | | Some | | None | | | |
| | Starting before enrolment | Starting on or after enrolment | Starting before enrolment | Starting on or after enrolment | | all | m | f |
| Group number | 1 | 2 | 3 | 4 | 5 | | | |
| Categorized rates per annum | | | | | | | | |
| 0 - none | 6 | 16 | 2 | 1 | 95 | 120 | 73 | 47 |
| <=1 | 12 | 28 | 5 | 10 | 99 | 154 | 81 | 73 |
| 1<2 | 8 | 32 | 8 | 20 | 74 | 142 | 73 | 69 |
| 2<3 | 14 | 26 | 7 | 18 | 65 | 130 | 67 | 63 |
| 3<4 | 9 | 37 | 5 | 30 | 60 | 141 | 77 | 64 |
| 4+ | 23 | 47 | 10 | 24 | 66 | 170 | 75 | 95 |
| N | 72 | 186 | 37 | 103 | 459 | 857 | 446 | 411 |

## Outcome: Body mass index – cohort over time

Cross-sectional analyses could be misleading as to the effect of ageing in individuals. The cohort enrolled in 1994 or 1995, who had been under observation for five or six years, was analysed for change in the BMI percentiles, standardized against the Child Growth Foundation 1990 references ("imputed" height calculations were used).

Although the groups were of mixed age, the composition of any particular group did not change, and for the initial analyses means by sex and dornase alfa use group were tabulated for inspection for any marked secular change (Table A.16). Of course, individuals are also ageing over time, and a separate analysis by sex and initial five-year age group is given in Table A.17, to see if there are any marked differences in the trends for different birth cohorts.

**Table A.16: Six-year cohort - Body Mass Index percentiles by sex, observation year, and dornase alfa groups, adjusted to 1990 Child Growth Foundation Data (n= 571)**

| | Use of dornase alfa | | | | | All | n - observations |
|---|---|---|---|---|---|---|---|
| | Continuous | | Some | | None | | |
| | Starting before enrolment | Starting on or after enrolment | Starting before enrolment | Starting on or after enrolment | | | |
| Group number | 1 | 2 | 3 | 4 | 5 | | |
| Sex m | 17 | 57 | 14 | 49 | 161 | 298 | |
| f | 22 | 81 | 13 | 37 | 120 | 273 | |
| Totals | 39 | 138 | 27 | 86 | 281 | 571 | |
| Males | | | | | | | |
| Year 1 | 30.0 | 27.0 | 41.9 | 28.4 | 45.2 | 37.9 | 297 |
| Year 2 | 28.3 | 28.6 | 44.1 | 30.2 | 45.5 | 38.7 | 297 |
| Year 3 | 33.4 | 29.0 | 43.4 | 31.6 | 44.1 | 38.6 | 293 |
| Year 4 | 31.4 | 29.1 | 39.0 | 30.2 | 43.8 | 37.9 | 295 |
| Year 5 | 27.8 | 29.8 | 36.2 | 31.2 | 44.2 | 38.0 | 292 |
| Year 6 | 33.0 | 33.1 | 31.2 | 36.2 | 43.5 | 39.2 | 216 |
| All years | 30.5 | 29.3 | 39.7 | 31.0 | 44.4 | 38.3 | 1690 |
| Females | | | | | | | |
| Year 1 | 26.9 | 34.6 | 35.8 | 36.2 | 46.6 | 39.5 | 269 |
| Year 2 | 27.2 | 32.7 | 34.9 | 37.5 | 48.3 | 39.9 | 272 |
| Year 3 | 26.2 | 31.4 | 34.2 | 40.6 | 47.1 | 39.3 | 273 |
| Year 4 | 26.0 | 29.9 | 34.4 | 38.9 | 47.3 | 38.7 | 262 |
| Year 5 | 21.8 | 29.3 | 31.1 | 41.4 | 46.6 | 38.0 | 270 |
| Year 6 | 24.8 | 25.3 | 40.5 | 37.8 | 42.4 | 35.5 | 185 |
| All years | 25.5 | 30.8 | 34.8 | 38.8 | 46.6 | 38.7 | 1531 |
| Males and females | | | | | | | |
| Year 1 | 28.3 | 31.5 | 39.0 | 31.7 | 45.8 | 38.7 | 566 |
| Year 2 | 27.7 | 31.0 | 39.6 | 33.3 | 46.7 | 39.2 | 569 |
| Year 3 | 29.2 | 30.4 | 38.8 | 35.5 | 45.4 | 38.9 | 566 |
| Year 4 | 28.4 | 29.5 | 36.8 | 33.8 | 45.3 | 38.2 | 557 |
| Year 5 | 24.4 | 29.5 | 33.8 | 35.6 | 45.2 | 38.0 | 562 |
| Year 6 | 28.7 | 28.9 | 35.4 | 36.9 | 43.0 | 37.5 | 401 |
| All years | 27.7 | 30.2 | 37.3 | 34.3 | 45.4 | 38.5 | 3221 |

**Table A.17: Six-year cohort - Body Mass Index percentiles by sex, observation year and initial age groups, adjusted to 1990 Child Growth Foundation Data (n= 571)**

| Age (years) | | 5<10 | 10<15 | 15<20 | 20<25 | 25<30 | 30<35 | 35+ | all | n - observations |
|---|---|---|---|---|---|---|---|---|---|---|
| Sex | m | 76 | 51 | 62 | 65 | 20 | 18 | 6 | 298 | |
| | f | 58 | 66 | 60 | 44 | 28 | 10 | 7 | 273 | |
| Totals | | 134 | 117 | 122 | 109 | 48 | 28 | 13 | 571 | |
| Males | | | | | | | | | | |
| Year 1 | | 52.7 | 42.8 | 35.0 | 30.8 | 18.5 | 20.6 | 35.4 | 37.9 | 297 |
| Year 2 | | 52.1 | 43.0 | 35.5 | 32.0 | 22.7 | 22.7 | 36.5 | 38.7 | 297 |
| Year 3 | | 51.3 | 42.3 | 33.6 | 33.8 | 24.4 | 24.3 | 39.2 | 38.6 | 293 |
| Year 4 | | 49.8 | 42.9 | 31.2 | 35.0 | 23.6 | 22.8 | 39.8 | 37.9 | 295 |
| Year 5 | | 48.5 | 44.1 | 28.7 | 38.8 | 26.5 | 20.4 | 33.1 | 38.0 | 292 |
| Year 6 | | 49.9 | 47.5 | 28.0 | 38.9 | 34.2 | 10.9 | 47.0 | 39.2 | 216 |
| All years | | 50.8 | 43.5 | 32.2 | 34.8 | 24.6 | 20.7 | 38.4 | 38.3 | 1690 |
| Females | | | | | | | | | | |
| Year 1 | | 48.6 | 37.0 | 44.2 | 34.2 | 31.6 | 32.4 | 25.3 | 39.5 | 269 |
| Year 2 | | 46.8 | 40.9 | 41.1 | 34.9 | 34.3 | 33.7 | 26.3 | 39.9 | 272 |
| Year 3 | | 43.9 | 41.4 | 39.6 | 32.8 | 36.3 | 39.0 | 30.6 | 39.3 | 273 |
| Year 4 | | 40.6 | 40.2 | 38.2 | 36.1 | 37.5 | 42.5 | 28.7 | 38.7 | 262 |
| Year 5 | | 37.2 | 41.6 | 34.5 | 38.2 | 37.9 | 46.8 | 29.3 | 38.0 | 270 |
| Year 6 | | 31.6 | 37.7 | 31.9 | 38.1 | 40.3 | 51.4 | 17.0 | 35.5 | 185 |
| All years | | 42.1 | 39.9 | 38.6 | 35.6 | 36.3 | 40.0 | 26.6 | 38.7 | 1531 |
| Males and females | | | | | | | | | | |
| Year 1 | | 51.0 | 39.6 | 39.5 | 32.2 | 26.2 | 24.8 | 29.9 | 38.7 | 566 |
| Year 2 | | 49.8 | 41.8 | 38.3 | 33.2 | 29.4 | 26.6 | 31.0 | 39.2 | 569 |
| Year 3 | | 48.1 | 41.8 | 36.6 | 33.4 | 31.3 | 29.5 | 34.6 | 38.9 | 566 |
| Year 4 | | 45.8 | 41.4 | 34.5 | 35.4 | 31.6 | 29.9 | 33.8 | 38.2 | 557 |
| Year 5 | | 43.7 | 42.6 | 31.5 | 38.6 | 33.2 | 29.9 | 30.9 | 38.0 | 562 |
| Year 6 | | 42.4 | 41.9 | 29.9 | 38.6 | 38.0 | 22.2 | 32.0 | 37.5 | 401 |
| All years | | 47.0 | 41.5 | 35.4 | 35.1 | 31.4 | 27.5 | 32.0 | 38.5 | 3221 |

# B Appendix B: Material relating to review of ESDs

- Table B.1: Data sources used in the ESDs, and number of patients in both
- Table B.2: Effectiveness outcomes in ESDs

## Table B.1: Data sources used in the ESDs, and number of patients in both

| Study reference | Year | Name of database(s) used | Purpose of database(s) | Number of patients in database | Number of patients included in analysis |
|---|---|---|---|---|---|
| (Benvegnu *et al.* 1998) | 1998 | Prospective follow-up study of cirrhosis patients since 1986 (unnamed) | Evaluating the course of the disease, development of hepato-cellular carcinoma, and death rate. | 290 | 189 |
| (Berman *et al.* 1997) | 1997 | Medicaid Management Information System | Claims database | 80143 | 12381 |
| (Bowman *et al.* 2000) | 2000 | Prescription database of PCS HealthSystems Inc. | Computerised prescription records of a pharmacy benefits management company (PCS HealthSystems Inc.) | >56 million | 58403 |
| (Butterworth *et al.* 1998) | 1998 | Benchmarking data within hospitals of the University Health System Consortium | Identifying and reducing unnecessary variations in processes of care for patients undergoing CABG | 1164 | 1094 |
| (Chew *et al.* 2001) | 2001 | Percutaneous coronary intervention registry at the Cleveland Clinic Foundation | Not stated | 885 | 860 |
| (Choi *et al.* 2002) | 2002 | Wichita Arthritis Center, National Death Index | Not stated | Not reported | 1240 |
| (Eggleston *et al.* 1996) | 1996 | Mediplus® UK | MediPlus provides anonymised clinical data mainly for market pharmaceutical research | Not reported | 279 |
| (Fedson *et al.* 1993) | 1993 | Manitoba population registry and Health Services Insurance Plan | Claims database | Not reported | 20137 (Double-counting likely) |
| (Gable *et al.* 1990) | 1990 | Blue Cross/Blue Shield of Minnesota | Claims database | Not reported | 762 |
| (Garcia Rodriguez & Ruigomez 1999) | 1999 | General Practice Research Database | Medical and health research | Not reported | 952 |
| (Ghani *et al.* 2001) | 2001 | Apache HIV Insight<br>Target Management Services<br>Clinical Partners | Commercial patient databases of clinics run by APACHE HIV Insight, Target Management Services, and Clinical Partners. | 14975 | 5597 |
| (Giralt *et al.* 2000) | 2000 | International Bone Marrow Transplant Registry | International database of clinical outcomes from now >450 blood and marrow transplant centres | Not reported | 873 |
| (Goldstein *et al.* 1996) | 1996 | Multicentre Study of Myocardial Ischemia | Measure the predictive capacity of non-invasive tests for subsequent coronary events | 936 | 936 |
| (Graham *et al.* 1991) | 1991 | Mulitcentre AIDS Cohort Study | Prospective study of natural history of HIV infection among men in the USA | 5494 | 2516 |

| Study reference | Year | Name of database(s) used | Purpose of database(s) | Number of patients in database | Number of patients included in analysis |
|---|---|---|---|---|---|
| (Heudebert *et al.* 1993) | 1993 | Database at a Veterans Affairs Medical Centre lipid clinic | Prospective patient database of a lipid clinic at a Veterans' Affairs medical centre | Not reported | 250 |
| (Huang *et al.* 1999) | 1999 | Medicaid paid claims data | Claims database | Not reported | 409 |
| (IBMTR 1989) | 1989 | International Bone Marrow Transplant Registry | Clinical outcomes database | Not reported | 634 |
| (Johnson *et al.* 1999a) | 1999 | The Epidemiologic Study of Cystic Fibrosis | Monitor patients receiving routine clinical care for CF | >21000 | 2665 |
| (Krumholz *et al.* 1995) | 1995 | Cooperative Cardiovascular Project pilot | Improve the quality of care for Medicare beneficiaries with AMI | 16189 | 10018 |
| (Krumholz *et al.* 1998) | 1998 | Cooperative Cardiovascular Project pilot | Improve the quality of care for Medicare beneficiaries with AMI | 16182 | 6935 |
| (Krumholz *et al.* 2001) | 2001 | Cooperative Cardiovascular Project | Examine patterns of care and stimulate improvements in the care and outcomes of Medicare beneficiaries with AMI | 234769 | 14129 |
| (Kuhn *et al.* 2000) | 2000 | Perinatal AIDS Collaborative Transmission Study (PACTS) | Not stated | Not reported | 325 |
| (Lawrenson & Logie 2001) | 2001 | General Practice Research Database | Medical and health research; GPRD currently collects data on ca. 3 million patients (5% of the UK population) | ca. 5.4 million | 75045 |
| (Lundgren *et al.* 1994) | 1994 | Aids in Europe Study Group | Not stated | 6578 | 4484 |
| (McDougall *et al.* 1994) | 1994 | Own clinical database of Rheumatic Disease Unit at Saskatoon (University of Saskatchewan) | Long-term prospective follow-up study of RA in one clinic | 1191 | 244 |
| (Moore *et al.* 1991) | 1991 | Human Immunodeficiency Virus Information System, National Death Index | Assess the volume, nature, and cost of health care services associated with HIV disease and related disorders | Not reported | 714 |
| (Nichol *et al.* 1994) | 1994 | Group Health, Inc. | Claims database (HMO) | >300,000 | 78527 (Double-counting likely) |
| (Nichol *et al.* 1999a) | 1999a | Group Health, Inc. | Claims database (HMO) | >250,000 | 1898 |
| (Nichol *et al.* 1999b) | 1999b | Group Health, Inc. | Claims database (HMO) | >250,000 | 1898 |

| Study reference | Year | Name of database(s) used | Purpose of database(s) | Number of patients in database | Number of patients included in analysis |
|---|---|---|---|---|---|
| (Nordin *et al.* 2001) | 2001 | HealthPartners<br>Oxford Health Plans<br>Kaiser Permanente Northwest | HMO databases | ca. 2.9 million | 281428 (Double-counting likely) |
| (Peterson *et al.* 1999) | 1999 | Second National Registry of Myocardial Infarction | Study of routine practice patterns and treatment outcomes (Ziegelstein *et al.*) | 351317 | 45925 |
| (Pethica *et al.* 1998) | 1998 | Database of the Royal New Zealand College of GPs Research Unit | Not stated | 128585 | 5930 |
| (Price *et al.* 1998) | 1998 | Thorpewood Primary Care Database | Computerised primary care practice database | Not reported | 37 |
| (Rabinowitz *et al.* 2001) | 2001 | a) National registry for novel antipsychotic use; b) National Psychiatric Hospitalization Case Registry | a) Enforce guidelines on use of novel antipsychotics; b) national registration of all psychiatric hospitalisations, including demographic data | Not reported | 1,039 |
| (Rahme *et al.* 2002) | 2002 | RAMQ = Regie de l'assurance maladie du Quebec<br>Med-Echo discharge summary database | RAMQ is insurance fund covering all aged 65 and over;<br>Med-Echo maintains a discharge summary database | Not reported | 28,323 |
| (Sebaldt *et al.* 1999) | 1999 | CANDOO - Canadian Database of Osteoporosis and Osteopenia | Prospective registry designed to capture comprehensive set of osteoporosis related clinical data during the course of routine specialist care | Not reported | 61 |
| (Sernyak *et al.* 2001) | 2001 | Veterans Affairs databases | Presumably claims database (VA) | Not reported | 3795 |
| (Solomon *et al.* 2002) | 2002 | Medicaid (New Jersey)<br>Medicare (New Jersey)<br>Pharmaceutical Assistance for the Aged and Disabled programs (New Jersey) | Claims databases | Not reported | 22125 |
| (Tiefenbrunn *et al.* 1998) | 1998 | Second National Registry of Myocardial Infarction | See Ziegelstein 2001 | 172742 | 28757 |
| (van Staa *et al.* 1998) | 1998 | General Practice Research Database | Medical and health research; | 3.5 million | 15954 |
| (Weintraub *et al.* 2001) | 2001 | Main source: Medicaid<br>Other sources: North Carolina dentist licensure surveys, North Carolina Statistical Abstracts, North Carolina Department of Health and Human Services | Claims database | Not reported | 15438 |
| (Ziegelstein *et al.* 2001) | 2001 | Second National Registry of Myocardial Infarction | Study of routine practice patterns and treatment outcomes (Ziegelstein *et al.*) | 173728 | 173728 |

## Table B.2: Effectiveness outcomes in ESDs

| First author | Year | Effectiveness outcomes | Category* |
|---|---|---|---|
| Benvegnu L. | 1998 | (1) Worsening of stage of cirrhosis (Child's classification), (2) development of HCC, (3) death or liver transplantation | C |
| Berman S. | 1997 | Rates of unresponsive acute otitis media; adverse drug reaction; costs, prescribing | C |
| Bowman L. | 2000 | Occurrence of another antibiotic treatment within 24 days of first dose;<br>average cost adjusted for age and sex | D |
| Butterworth J. | 1998 | Duration of intubation, length of stay | C |
| Chew D. | 2001 | Death, myocardial infarction | M |
| Choi H. | 2002 | Mortality; cardiovascular and non-cardiovascular mortality | M |
| Eggleston A. | 1996 | Success rates = reflux medication prescribed for <=3 months; relapse rates in 6 months in successfully treated patients. | D |
| Fedson D. | 1993 | Discharge with pneumonia and influenza and all respiratory conditions; hospital deaths with same; all deaths due to respiratory diseases and all causes | C |
| Gable C. | 1990 | Pneumonia incidence rate ratio of pre- and post-vaccination periods in vaccinated and non-vaccinated groups | C |
| García L. | 1999 | Upper gastrointestinal bleeding | C |
| Ghani A. | 2001 | Time taken for HIV-1 RNA to fall below detectable levels | C |
| Giralt S. | 2000 | Leukaemia-free survival (i.e. in complete remission) | C |
| Goldstein R. | 1996 | Cardiac mortality;<br>Non-fatal myocardial infarction, unstable angina | M |
| Graham N. | 1991 | Progression to AIDS | C |
| Heudebert G. | 1993 | Percentage difference between the mean basal LDL-C level and the mean of all LDL-C measurements for the same patient during drug therapy | C |
| Huang X. | 1999 | "Cure" rate (i.e. no further treatment with one of three drugs within 30 days after the initial treatment) | D |
| IBMTR | 1989 | Relapse (i.e. reappearance of medullary or extramedullary leukaemia) | C |
| Johnson C. | 1999 | Change of $FEV_1$ (%predicted) from baseline to 12-month follow-up | C |
| Krumholz H. | 1998 | 30-day mortality;<br>in-hospital haemorrhage, in-hospital transfusion, in-hospital stroke, 1-year mortality. | M |
| Krumholz H. | 1995 | 30-day mortality;<br>use of aspirin within first 2 hospital days | M |
| Krumholz H. | 2001 | Mortality within 1 year of discharge | M |
| Kuhn L. | 2000 | Death or diagnosis with a category C disease within the first year of life | M |
| Lawrenson R. | 2001 | Treatment failure (further antibiotic within 28 days) | D |
| Lundgren J | 1994 | Mortality;<br>use of zidovudine | M |
| McDougall R. | 1994 | Disease activity, functional class, mortality | C |
| Moore R. | 1991 | Survival | M |
| Nichol K. | 1994 | Rate of hospitalisation for influenza and complications of influenza, including pneumonia, all acute and chronic respiratory conditions, and congestive heart failure, and its effect on mortality; | C |

| | | | |
|---|---|---|---|
| | | costs of hospitalisation | |
| Nichol K. | 1999a | Hospitalisations for pneumonia and influenza, for all respiratory conditions; and death from all causes;<br>outpatient visits for pneumonia and influenza and all respiratory conditions | C |
| Nichol K. | 1999b | Hospitalisations for pneumonia (including influenza);<br>death from all causes, hospitalisation costs | C |
| Nordin J. | 2001 | Hospitalisations for pneumonia and influenza;<br>deaths from all causes | C |
| Peterson L. | 1999 | In-hospital mortality;<br>In-hospital mortality + nonfatal stroke | M |
| Pethica B. | 1998 | Mean prescribed daily inhaled corticosteroid dose | D |
| Price D. | 1998 | Peak expiratory flow, drug requirements, consultations and attendances | C |
| Rabinowitz J. | 2001 | Rehospitalisation | C |
| Rahme E. | 2002 | AMI discharge diagnosis | C |
| Sebaldt R. | 1999 | Changes within groups from baseline and differences between groups in bone mineral density of lumbar spine, femoral neck, and trochanter | C |
| Sernyak M. | 2001 | Psychiatric inpatient days | C |
| Solomon D. | 2002 | AMI | C |
| Tiefenbrunn A. | 1998 | In-hospital mortality;<br>In-hospital mortality + nonfatal stroke | M |
| van Staa T. | 1998 | Number of patients with a fracture (not number of fractures!) | C |
| Weintraub J. | 2001 | Caries-related services involving the occlusal surface of permanent first molars (CRSOs);<br>cumulative expenditure | C |
| Ziegelstein RC. | 2001 | Mortality | M |

* C=clinical outcome; D=drug-related outcome; M=mortality outcome

# C Appendix C: Material relating to comparison case studies

a. Items assessed by quality assessment instruments and checklists for non-randomised studies

b. Quality Assessment Schedule for ESDs

c. Guide to Quality Assessment Schedule for ESDs

d. Quality of ESDs used for case studies

e. Data extraction sheet for review cases studies

## Table C.1: Items assessed by quality assessment instruments and checklists for non-randomised studies

For CRD 4 and Newcastle/Ottawa:
Times New Roman - suggestions for cohort studies
Arial Narrow – suggestions for case-control studies
**Bold** - suggestions for both cohort and case control studies
Bradley Hand - suggestions for case series (CRD 4 only)

| Zaza et al. (2000) | Centre for Reviews and Dissemination Version 4 (NHS Centre for Reviews and Dissemination 2001) | Newcastle-Ottawa Scale (Wells et al. 2003) | Downs & Black (1997) with amendments and suggestions by MacLehose et al. (2000) |
|---|---|---|---|
| Description of population | Description of groups | | Patient characteristics clearly described |
| Description of intervention | | | Interventions / exposure of interest clearly described |
| Sampling - frame | Representative sample of relevant population? | Representativeness of cases? | Proportion ineligible |
| | | | Proportion refusing to participate |
| | Controls randomly selected from source of population of cases? | Selection of controls | |
| | | | Staff, places, and facilities representative of majority of patients? |
| | | Representativeness of exposed cohort | |
| | | | Inclusion / exclusion criteria clearly described. |
| | | Selection of non-exposed group | |
| | | | Patients in different groups from same population? |
| Screening criteria for eligibility | Case definition explicit | Case definition adequate? | Inclusion / exclusion criteria clearly stated |
| | Criteria for inclusion explicit? | | |
| | | Definition of controls | |
| Unit of analysis population entire population or probability sample | Proportion followed up | | |
| Selection bias issues | | Demonstration that outcome not present at start of study | |
| Exposure variables - valid | | Method and blinding of exposure | Measurement of intervention / exposure valid? |

| Zaza et al. (2000) | Centre for Reviews and Dissemination Version 4 (NHS Centre for Reviews and Dissemination 2001) | Newcastle-Ottawa Scale (Wells et al. 2003) | Downs & Black (1997) with amendments and suggestions by MacLehose et al. (2000) |
|---|---|---|---|
| Exposure variables - reliable | Intervention / treatment reliably ascertained?<br><br>Interventions and exposures assessed in the same way across groups<br><br>Disease state of cases reliably assessed and validated | | |
| Outcome and other independent variables - valid | Outcome assessment blind to exposure status?<br><br>Outcomes assessed by objective criteria or blinding used?<br><br>Description of distribution of prognostic factors | Assessment of outcome - method and blinding | Attempt made to blind patients?<br><br>Attempt made to blind assessors?<br><br>Attempt made to blind those performing intervention? |
| Outcome and other independent variables - reliable | | | Outcome measures valid and reliable?<br><br>Main confounding variables valid and reliable? |
| Analysis | Adequate adjustment for effect of confounding factors<br><br>Appropriate statistical analysis used? (matched or unmatched)<br><br>If comparisons of sub-series, was description of series and distribution of prognostic factors sufficient?<br><br>Is it possible that over-matching has occurred and factors related to exposure were | Comparability of cases and controls (cohorts) on the basis of the design or analysis | 95% CI and/or actual probability values for main outcomes reported?<br><br>Estimates of random variability of main outcomes (7).<br><br>If any results based on "data dredging" - is this made clear?<br><br>Statistical tests legitimate?<br><br>Adequate adjustment for confounding in analyses?<br><br>Intention to treat? |

| Zaza et al. (2000) | Centre for Reviews and Dissemination Version 4 (NHS Centre for Reviews and Dissemination 2001) | Newcastle-Ottawa Scale (Wells et al. 2003) | Downs & Black (1997) with amendments and suggestions by MacLehose et al. (2000) |
|---|---|---|---|
| | matched for? | | |
| Interpretation of results - % completed | Dropout rates and reasons similar across groups? | Non-response rate in both groups | Numbers lost to follow-up described? |
| | | | Proportion in each group completing treatment? |
| | Response rates and reasons similar across groups? | | How many subjects lost to follow-up? |
| | | | Comparisons made, in case of substantial losses to follow-up? |
| Interpretation of results - comparability of groups | Groups comparable on all important confounding factors | | Distribution of confounders clearly described for each group |
| | Comparable with respect to confounding factors? | | Patients in different groups recruited over same period of time? |
| Interpretation of results - control to limit bias | Individuals entered at similar point in their disease progression? | | Patients randomised? |
| Biases addressed and un-addressed | How was response rate defined? | Follow-up long enough for outcomes to occur? | Effects of patients' preferences and expectations on outcomes considered? |
| | Follow-up long enough for outcomes to occur? | Adequacy of follow-up | |
| | | | Any planned analyses subject to bias? |
| | Follow-up long enough for events to occur? | | Follow-up duration same in all groups? |
| | Dose-response relationship demonstrated | | |
| Reporting | | | Clear description of: Hypothesis/aim/objective, main outcome, main findings |
| | | | Interventions appropriate? |
| | | | Power calculation reported? |
| | | | Primary outcome identified? |
| | | | All possible adverse events reported? |

# QUALITY ASSESSMENT SCHEDULE FOR ESDs
## Version 18 December 2002

**TO BE USED WITH ACCOMPANYING GUIDE.**

**I. INTERNAL VALIDITY**

| | yes | no | unsure | N/a | Source* |
|---|---|---|---|---|---|
| 1) Was the exposure / intervention variable: Valid | ☐ | ☐ | ☐ | ☐ | Z-m |
| Reliable (consistent and reproducible)? | ☐ | ☐ | ☐ | ☐ | |
| 2) Were the outcome and other independent (or predictor) variables: Valid | ☐ | ☐ | ☐ | ☐ | Z |
| Reliable (consistent and reproducible)? | ☐ | ☐ | ☐ | ☐ | |
| 3) Were outcome assessors aware of the study objectives? | ☐ | ☐ | ☐ | ☐ | |
| 4) Were those administering or prescribing treatment aware of the study objectives? | ☐ | ☐ | ☐ | ☐ | |
| 5) In case-control studies, has the disease state of cases been reliably assessed and validated? | ☐ | ☐ | ☐ | ☐ | |

| | Yes | No-accounted for | No, not accounted for | Source* |
|---|---|---|---|---|
| 6) Was the duration of follow-up the same for all groups being compared? | ☐ | ☐ | ☐ | Mac-m |

| | yes | no | unsure | N/a | Source* |
|---|---|---|---|---|---|
| 7) Proportion of patients in intervention group lost to follow-up (prospective study) or excluded from analysis (retrospective study) is <5%. | ☐ | ☐ | ☐ | ☐ | |
| 8) Proportion of patients in control group lost to follow-up (prospective study) or excluded from analysis (retrospective study) is <5%. | ☐ | ☐ | ☐ | ☐ | |
| 9) If substantial losses to follow-up occurred (i.e. >5%) was a comparison made of the characteristics of those lost to follow-up and those followed up? | ☐ | ☐ | ☐ | ☐ | Mac-m |
| 10) Are patients in the study group naïve to the drug or free of use for a significant timespan before the treatment period in question? | ☐ | ☐ | ☐ | ☐ | |
| 11) Did the authors assess whether the units of analyses were comparable prior to exposure to the intervention? | ☐ | ☐ | ☐ | ☐ | Z |
| 12) Did the authors correct for controllable variables or institute study procedures to limit bias appropriately (e.g., randomisation, restriction, matching, stratification, or statistical adjustment)? | ☐ | ☐ | ☐ | ☐ | Z-m |
| 13) Did the authors identify and discuss potential biases or unmeasured/contextual confounders that may account for or influence the observed results? | ☐ | ☐ | ☐ | ☐ | Z-m |

14) If yes, did they explicitly state how they assessed these potential confounders and biases? ☐ ☐ ☐ ☐ Z-m

15) List biases / confounders <u>not</u> identified and discussed by authors:

**II. EXTERNAL VALIDITY**

| | yes | no | unsure | N/a | |
|---|---|---|---|---|---|
| 1) Did the authors clearly describe the data source from where the study population was drawn? | ☐ | ☐ | ☐ | ☐ | Z-m |
| 2) Did the authors clearly define the screening criteria for study eligibility? | ☐ | ☐ | ☐ | ☐ | Z-m |
| 3) What proportion of subjects in the database were ineligible to participate? | ☐ | | | % | Mac-m |
| 4) What proportion of subjects who were eligible were excluded from the analysis? | ☐ | | | % | Mac-m |
| 5) Were the staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients receive? | (1) ☐ ☐ ☐ ☐ ☐ | | | (0) | Mac-m |
| | yes | no | unsure | N/a | |
| 6) Is it likely that patients also receive relevant care from alternative sources? | ☐ | ☐ | ☐ | ☐ | (MF) |
| 7) Was the population that served as the unit of analysis the entire eligible population or a probability sample at the point of observations? | ☐ | ☐ | ☐ | ☐ | Z |
| 8) Were the patients in different intervention groups (trials and cohort studies) or the cases and controls (case-control studies) recruited from the same population? | ☐ | ☐ | ☐ | ☐ | Mac |
| 9) Were the patients in the different treatment group (trials and cohort studies) or the cases and controls (case-control studies) recruited over the same period of time? | ☐ | ☐ | ☐ | ☐ | Mac -n |
| 10) Are there other selection bias issues not identified above? Describe: | ☐ | ☐ | ☐ | ☐ | Z |

**III. DATA PROCESSING**

Access and Confidentiality

| | yes | no | unsure | N/a |
|---|---|---|---|---|
| 1) Is there any indication of any of the authors being involved with the running of the database or registry? | ☐ | ☐ | ☐ | ☐ |
| 2) Is there any indication that a Research Ethics Committee or Institutional Review Board or similar body has approved the research study? | ☐ | ☐ | ☐ | ☐ |
| 3) Is there any indication that patients have given informed consent to the research study? | ☐ | ☐ | ☐ | ☐ |

Source*

4) Are individuals' data records identifiable through a code or record ID number (regardless of whether the researchers can decode the information or not)? ☐ ☐ ☐ ☐

5) Are any safeguards on confidentiality or data protection reported? ☐ ☐ ☐ ☐

Routine data collection and processing procedures

| | yes | no | unsure | N/a |
|---|---|---|---|---|
| 6) Is any routine data validation of variables reported? | ☐ | ☐ | ☐ | ☐ |
| 7) Has any verification against alternative sources of data been reported on any data in the database or registry (either for the current or previous studies)? | ☐ | ☐ | ☐ | ☐ |
| 8) Does or did routine processing include manual copying / entry of data? | ☐ | ☐ | ☐ | ☐ |
| 9) Is the aim of the database or registry congruous with the aim of the study? | ☐ | ☐ | ☐ | ☐ |

10) Number of centres involved in the database / registry: ☐ Not reported: ☐

11) Are centre variations in clinical or reporting practices likely? ☐ ☐ ☐ ☐

Assessment of data quality including validation

| | yes | no | unsure | N/a |
|---|---|---|---|---|
| 12) Does the paper report that the included data have been assessed for consistency, accuracy, and completeness (either by database operators or researchers)? | ☐ | ☐ | ☐ | ☐ |
| 13) Do the researchers report having undertaken any assessment of data quality? | ☐ | ☐ | ☐ | ☐ |

Individual variables

| | yes | no | unsure | N/a |
|---|---|---|---|---|
| 14) Are the inclusion and exclusion criteria directly measured in the database or registry or any linked data source? | ☐ | ☐ | ☐ | ☐ |
| 15) Are the main outcome and exposure / intervention of interest directly measured in the database or registry or any linked data source? | ☐ | ☐ | ☐ | ☐ |
| 16) Are the identified predictors and confounders directly measured in the database or registry or any linked data source? | ☐ | ☐ | ☐ | ☐ |
| 17) If "no" to any of the previous three questions, have all used algorithms been validated? | ☐ | ☐ | ☐ | ☐ |

**IV. ANALYSIS**

1) Are estimates of the random variability of the main outcomes clearly described for each group of patients to be compared? ☐ ☐ ☐ ☐ Mac-m

| | yes | no | unsure | N/a | Source |
|---|---|---|---|---|---|
| 2) Did the authors conduct appropriate statistical testing by: | ☐ | ☐ | ☐ | ☐ | Z |
| i) Conducting statistical testing (when appropriate)? | ☐ | ☐ | ☐ | ☐ | Z |
| ii) Reporting which statistical tests were used? | ☐ | ☐ | ☐ | ☐ | Z |
| ii) Controlling for design effects in the statistical model? | ☐ | ☐ | ☐ | ☐ | Z |
| iv) Controlling for repeated measures in populations that were followed over time? | ☐ | ☐ | ☐ | ☐ | Z |
| v) Controlling for differential exposure to the intervention? | ☐ | ☐ | ☐ | ☐ | Z |
| vi) Using a model designed to handle multi-level data when they included group-level and individual covariates in the model? | ☐ | ☐ | ☐ | ☐ | Z |
| vii) Allowing for important centre differences in the analysis? | ☐ | ☐ | ☐ | ☐ | |

Missing or incomplete data

| | yes | no | unsure | N/a | |
|---|---|---|---|---|---|
| 3) Have patients with incomplete data been excluded? | ☐ | ☐ | ☐ | ☐ | |
| 4) Have statistical adjustments been used for missing data? | ☐ | ☐ | ☐ | ☐ | |

Multilevel data

| | yes | no | unsure | N/a | |
|---|---|---|---|---|---|
| 5) Have any measures of the overall adequacy of the statistical model been reported? | ☐ | ☐ | ☐ | ☐ | (MF) |
| 6) Have any sensitivity analyses been performed? | ☐ | ☐ | ☐ | ☐ | (MF) |

**V. REPORTING**

| | yes | no | unsure | N/a | Source * |
|---|---|---|---|---|---|
| 1) Was the study population well described? | ☐ | ☐ | ☐ | ☐ | Z-m |
| 2) Was the intervention well described (what, how, who, where)? | ☐ | ☐ | ☐ | ☐ | Z-m |
| 3) Have the numbers of patients lost to follow-up been described? | ☐ | ☐ | ☐ | ☐ | Mac |
| 4) Was a power calculation reported or at least mentioned? | ☐ | ☐ | ☐ | ☐ | |
| 5) Are the main findings of the study clearly described? | ☐ | ☐ | ☐ | ☐ | Mac-m |
| 6) Have any explicit considerations of alternative explanations of the findings been presented? | ☐ | ☐ | ☐ | ☐ | (MF) |
| 7) Are conclusions going beyond capabilities of the database or study design? | ☐ | ☐ | ☐ | ☐ | (MF) |

**(MF) = question based on ideas by Motheral & Fairman 1997, Z = question from Zaza 2000, Mac = question from MacLehose 2000 (based on Downs & Black 1997), m = modified*

# GUIDE TO QUALITY ASSESSMENT SCHEDULE FOR ESDs Version 18 December 2002

## I. INTERNAL VALIDITY

1) Were the exposure / intervention variables valid measures of the intervention under study?
**The author should have reported one or more of the following:**
Clear definition of the exposure variable.
Measurement of exposure in different ways. ***Example:*** *consistency checks for self reports; use of corroborating respondents; programme or organisational record searches compared to self-reports.*
Citations or discussion as to why the use of these measures is valid. ***Example:*** *the authors considered evidence from similar studies, or available standards of measurement.*
Other

Were the exposure / intervention variables reliable (consistent and reproducible) measures of the intervention under study?
**The author should have reported one or more of the following:**
**Measures of internal consistency.** *Example:* ***Cronbach's alpha; confirmatory factor analysis.***
Measurement of exposure in different ways. ***Example:*** *see above.*
Inter-rater reliability checks (if exposure was determined by an observer.) ***Example:*** *percent agreement, kappa*
Citations or discussions as to why the use of these measures is reliable. ***Example:*** *see above.*
Other.

2) Were the outcome and other independent (or predictor) variables valid measures of the outcome of interest?
**The author should have reported one or more of the following:**
Clear definition of the outcome variable.
Measurement of the outcome in different ways. ***Example:*** *Correlational analysis between measured outcomes to demonstrate convergent (i.e., 2 or more measures reflect the same underlying process) or divergent validity (i.e., 2 or more measures reflect different dimensions). An example of the former is that 5 items on self-efficacy correlate highly with each other; an example of the latter is that self-efficacy measures do not correlate highly with attitude measures.*
Citations or discussion as to why the use of these measures is valid. ***Example:*** *see above.*
Other. ***Example:*** *If authors fail to blind observers/interviewers to treatment vs. comparison group, when applicable, the answer to this question should be "no".*

Were the outcome and other independent (or predictor) variables reliable (consistent and reproducible) measures of the outcome of interest?
**The author should have reported one or more of the following:**
Measures of internal consistency. ***Example:*** *see above.*
Measurement of outcome in different ways. ***Example:*** *see above.*
Considered consistency of coding, scoring or categorisation between observers (e.g., inter-rater reliability checks) or between different outcome measures. ***Example:*** *percent agreement, kappa*
Considered how setting and sampling of study population might affect reliability.
Citations or discussion as to why the use of these measures is reliable. ***Example:*** *see above.*

3) *Were outcome assessors aware of the study objectives?*

4) *Were those administering or prescribing treatment aware of the study objectives?*

5) *Was the duration of follow up the same for all groups being compared?*

6) *Proportion of patients in intervention group lost to follow-up (prospective study) or excluded from analysis (retrospective study) is <5%.*

7) *Proportion of patients in control group lost to follow-up (prospective study) or excluded from analysis (retrospective study) is <5%.*

8) *If substantial losses to follow-up occurred (i.e. >5%) was a comparison made of the characteristics of those lost to follow-up and those followed up?*
For retrospective studies, interpret "lost to follow-up" as meaning refusal by eligible cases/intervention subjects or controls to take part.

9) *Are patients in the study group naïve to the drug or free of use for a significant timespan before the treatment period in question?*

10) *Did the authors assess whether the units of analyses were comparable prior to exposure to the intervention?* For example, they should have assessed likely confounding via report of p- values and confidence intervals for the descriptive variables of age and sex or other key individual/community characteristics.

11) *Did the authors correct for controllable variables or institute study procedures to limit bias appropriately (e.g., randomisation, restriction, matching, stratification, or statistical adjustment)?*
Considering the study design, were appropriate methods for controlling confounding variables and limiting potential biases used? Confounding can be addressed by appropriate use of randomisation, restriction, matching, stratification, or multivariable methods. Sometimes use of a single method may be inadequate. Some biases con be limited by institution or data collection or study procedures that support validity of the study (e.g. training and/or blinding of interviewers or observers, interviewers and observers are different from intervention implementers etc.)
***Example:*** *If between-group differences persist after randomisation or matching, statistical control should also have been used.*

12) *Did the authors identify and discuss potential biases or unmeasured/contextual confounders that may account for or influence the observed results?*

13) *If yes, did they explicitly state how they assessed these potential confounders and biases?*

14) *List biases / confounders <u>not</u> identified and discussed by authors:*

## II. EXTERNAL VALIDITY

### Representativeness and Applicability

1) *Did the authors clearly describe the data source from where the study population was drawn?*
A clear description of the data source should include its size, characteristics of included population, geographic area, and time period covered.

2) *Did the authors specify the screening criteria for study eligibility* (if applicable)?

3) *What proportion of subjects in the database were ineligible to participate?*
"Ineligible" are those who do not meet the inclusion criteria. Answer 0% for case-control studies where the cases and controls would not be approached if they were ineligible.

4) *What proportion of subjects who were eligible were excluded from the analysis?*
For case-control studies answer 0%. The refusal to participate in case-control studies gives rise to problems of selection bias, not external validity/generalisability.

5) *Were the staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients receive?*
For the question to be answered "representative" the reader should be confident that the findings of the study would apply in a range of different settings (e.g. teaching hospitals and district general hospital). Score on a scale between 1 (box furthest to the left) and 0, with 1 being representative, 0 being not representative.

6) *Is it likely that patients also receive relevant care from alternative sources?*
Relevant care here means care or treatment similar to the intervention or which may significantly affect the outcome.

7) *Was the population that served as the unit of analysis the entire eligible population or a probability sample at the point of observations?*
If the sample contains the entire eligible population or a probability sample, answer "yes" otherwise "no".

8) *Were the patients in different intervention groups (trials and cohort studies) or the cases and controls (case-control studies) recruited from the same population?*
For example, patients for all comparison groups should be selected from the same hospitals. The question should be answered "can't tell" for cohort studies where there is no information concerning the source of patients included in the study. It should be considered that for some case-control studies that the use of the same local population for cases and controls may be inappropriate. A subjective judgement has to be made in the instance of case-control studies, as on occasion the use of the same local population for controls as well as cases may be inappropriate.

9) *Were the patients in the different treatment group (trials and cohort studies) or the cases and controls (case-control studies) recruited over the same period of time?*
For a study which does not specify the time period over which patients were recruited, the question should be answered "unsure". Also, if the data collection period stretches over several years, consider whether this might have an impact on period or cohort effects.

10) *Are there other selection bias issues not identified above?*
This might include a very low participation rate (or a high refulsal rate), an all-volunteer sample (as opposed to a convenience sample selected by the investigators), an inappropirate control or comparison group, or extremely restricted sampling inappropriate for measuring the effectiveness of the intervention being studied.

## III. DATA PROCESSING

### Access and Confidentiality

1) Is *there any indication of any of the authors being involved with the running of the database or registry?*
If there is no indication of any organisational or personal involvement, answer "no". Such indications may appear in the addresses and affiliations of authors or references to the database or registry. If operations of the database are narrated in the first person ("we"), assume that there was involvement.

2) *Is there any indication that a Research Ethics Committee or Institutional Review Board or similar body has approved the research study?*
In case of GPRD studies answer "yes"; these studies are approved by a specific ethics committee.

3) Is *there any indication that patients have given informed consent to the research study?*
If consent is mentioned, answer "yes".

4) *Are individuals' data records identifiable through a code or record ID number (regardless of whether the researchers can decode the information or not)?*

For longitudinal data sources, answer "yes".

5) Are *any safeguards on confidentiality or data protection reported?*
Examples are any references to measures in order to comply with relevant regulations, references to data security and protection of patient identity.

**Routine data collection and processing procedures**

6) *Is any routine data validation of variables reported?*
Any checking procedures applied routinely to the data, e.g. checks for consistency, completeness, etc.

7) *Is any routine testing of reliability of measurments reported?*

8) *Has any verification against alternative sources of data been reported on any data in the database or registry (either for the current or previous studies)?*
Alternative sources of data may be other databases, patient records, etc.

9) *Does or did routine processing include manual copying / entry of data?*
Routine processing here means the procedures involved in gathering and maintaining the data for/in the database or registry, not extracting or processing them for the study.

10) *Number of centres involved in the database / registry:*

10) *Are centre variations in clinical or reporting practices likely?*
Answer "yes", if there is no comment on standardised reporting procedures or if treatment choice is not influenced centrally but entirely dependent on local clinicians.
Specify which differences may exist.

**Assessment of data quality including validation**

11) *Does the paper report that the included data have been assessed for consistency, accuracy, and completeness (either by database operators or researchers)?*

12) *Do the researchers report having undertaken any assessment of data quality?*

**Individual variables**

14) - 16)
Directly measured means not having been derived through assumptions involving other variables, e.g. exacerbations have been recorded as such, rather than derived from temperature readings and culture tests.

17) *If "no" to any of the previous three questions, have all used algorithms been validated?*
Testing of algorithms may involve validation against other sources of information.

**IV. ANALYSIS**

1) *Are estimates of the random variability of the main outcomes clearly described for each group of patients to be compared?*
In non-parametric data the inter-quartile range of results should be reported. In parametric data

the standard error, standard deviation, or confidence intervals should be reported. For binary outcomes, confidence intervals should be reported for each group. For case-control studies, confidence intervals should be reported for the proportions of cases and controls who are exposed.

2) Check "yes", "no", or "not applicable" for each of the following:
Did the authors conduct appropriate analysis by:
Conducting statistical testing (when appropriate)?
Reporting which statistical tests were used?
Controlling for design effects in the statistical model?
*Examples:*
*The study population was sampled using complex stratified sampling, however, the authors did not control For the sampling method in the analysis.*
*The answer should be "no" if the study had a matched design but an unmatched analysis.*
Controlling for repeated measures in the analysis, for study designs in which the same population was followed with repeated measurements over time?
Accounting for different levels of exposure in segments of the study population in the analysis?
If the authors analysed group-level and individual-level covariates in the same statistical model, was the model designed to handle multi-level data?
Where centre differences may impinge on analysed variables, did the analysis use multi-level methods?

**Missing or incomplete data**

3) *Have patients with incomplete data been excluded?*

4) *Have statistical adjustments been used for missing data?*

**Multi-level data**

5) *Have any measures of the overall adequacy of the statistical model been reported* (e.g. $R^2$ in multiple regression analyses)?

6) *Have any sensitivity analyses been performed?*
Sensitivity analyses of impact of methodological decisions such as criteria for selection of patients, treatment patterns, or diagnosis proxy variables.

## V. REPORTING

1) *Was the study population* (i.e. the intervention and comparison population) *well described?*
The study population should be described by time (e.g., when the study population received the intervention), place, and person. Information about "person" should include at least age (for all studies) and should include other relevant characteristics of participants that are key to a particular study (e.g., SES, gender, other). Important potential confounding factors should also be described.

2) *Was the intervention well described?*
The intervention should be described in terms of what was done, how it was delivered, who was targeted, and where it was done.

3) *Have the numbers of patients lost to follow-up been described?*
Loss to follow-up should be interpreted as "refusal to take part by eligible subjects" for case-control and retrospective cohort studies.

4) *Was a power calculation reported or at least mentioned?*

5) *Are the main findings of the study clearly described?*
Simple outcome data (including denominators and numerators) should be reported for all major

findings so that the reader can check the major analyses and conclusions. (This question does not cover statistical tests, which are considered above).

6) *Have any explicit considerations of alternative explanations of the findings been presented?*

7) *Are conclusions going beyond capabilities of the database or study design?*

**Quality of ESDs included as case studies**

The quality assessment form for ESDs could not be used on Eggleston *et al.* 1996, because the published abstract contained very limited information on which to base the assessment. Specific issues which may impinge on the validity of this study, beyond those common to ESDs, are:

1. There is no indication of how far back records have been checked for previous diagnoses of GORD;
2. Random sampling has been undertaken *before* study eligibility has been assessed; this restricts the sample size unnecessarily;
3. Outcome assessment is based on GORD treatment variables (assessment of "relapse" after 6 months is not defined).

The results of the quality assessment of the other ESDs are presented in Table C.2 below.

## Table C.2: Quality assessment of ESDs (except Eggleston)

| | McDOUGALL | SEBALDT | TIEFEN-BRUNN | VAN STAA |
|---|---|---|---|---|
| **I. INTERNAL VALIDITY** | | | | |
| 1) Was the exposure / intervention variable: | | | | |
| Valid | No | No | No | Possibl |
| Reliable (consistent and reproducible)? | No | No | No | No |
| 2) Were the outcome and other independent (or predictor) variables: | | | | |
| Valid | No | No | Some | Yes |
| Reliable (consistent and reproducible)? | No | Yes, calibrated regularly | No | No |
| 3) Were outcome assessors aware of the study objectives? | No | Unsure | Probably not | No |
| 4) Were those administering or prescribing treatment aware of the study objectives? | No | Unsure | No | No |
| 5) In case-control studies, has the disease state of cases been reliably assessed and validated? | N/a | N/a | N/a | N/a |
| 6) Was the duration of follow-up the same for all groups being compared? | No? | Yes | Yes | Yes |
| 7) Proportion of patients in intervention group lost to follow-up (prospective study) or excluded from analysis (retrospective study) is <5%. | No | Yes | No | Yes |
| 8) Proportion of patients in control group lost to follow-up (prospective study) or excluded from analysis (retrospective study) is <5%. | No | (Complete follow-up only for one of the outcomes) | No | Yes |
| 9) If substantial losses occurred (i.e. >5%) was a comparison made of the characteristics of those lost to follow-up and those followed up? | N/a | N/a | N/a | No |
| 10) Are patients in the study group naïve to the drug or free of use for a significant time span before the treatment period in question? | Yes | Not reported, but most likely | Not reported, but most likely | Unsure |
| 11) Did the authors assess whether the units of analyses were comparable prior to exposure to the intervention? | Yes | Yes | Yes | Yes |
| 12) Did the authors correct for controllable variables or institute study procedures to limit bias appropriately (e.g., randomisation, restriction, matching, stratification, or statistical adjustment)? | Yes | No (BMD baseline differences not adjusted for) | Yes | Yes |
| 13) Did the authors identify and discuss potential biases or unmeasured/contextual confounders that may account for or influence the observed results? | Yes | No | Yes | Yes |
| 14) If yes, did they explicitly state how they assessed these potential confounders and biases? | Yes | N/a | Yes | Yes |
| 15) List biases / confounders <u>not</u> identified and discussed by authors: | N/a | Gender differences not tested; attrition in | Centre difference s not accounted | Dose an duration of treatme |

| | McDOUGALL | SEBALDT | TIEFEN-BRUNN | VAN STAA |
|---|---|---|---|---|
| | | controls not explained or tested; single-centre study - not clear who assessed outcomes and knew about study objectives. | for; | t not taken into account |
| **II. EXTERNAL VALIDITY** | | | | |
| Representativeness and Applicability | | | | |
| 1) Did the authors clearly describe the data source from where the study population was drawn? | Yes | No, but there is a reference | No, but there is a reference | Yes |
| 2) Did the authors clearly define the screening criteria for study eligibility? | No | No: exclusion criteria not clear (e.g. which drugs were permitted?); unclear what is "ambulatory" | Yes | Yes |
| 3) What proportion of subjects in the database was ineligible to participate? | Unclear | Unknown | Unknown | 99.5% |
| 4) What proportion of subjects who were eligible was excluded from the analysis? | Unclear | Apparently 0% | Apparently 0% | All etidronate takers should be included |
| 5) Were the staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients receive? (0 - 0.25 - 0.5 - 0.75 - 1) | 0.5 | Unknown | Unknown | Yes |
| 6) Is it likely that patients also receive relevant care from alternative sources? | Unknown | Unknown | No (inpatient data only) | Unknown |
| 7) Was the population that served as the unit of analysis the entire eligible population or a probability sample at the point of observations? | Unclear | Yes | Yes | Yes |
| 8) Were the patients in different intervention groups (trials and cohort studies) or the cases and controls (case-control studies) recruited from the same population? | Yes | Yes | Yes | Yes |
| 9) Were the patients in the different treatment group (trials and cohort studies) or the cases and controls (case-control studies) recruited over the same period of time? | No | Unsure | Yes | Yes |
| 10) Are there other selection bias issues not otherwise addressed? | N/a | Timing is not addressed; control patients seem to be | N/a | N/a |

| | McDOUGALL | SEBALDT | TIEFEN-BRUNN | VAN STAA |
|---|---|---|---|---|
| | | largely refusers of etidronate | | |
| **III. DATA PROCESSING** | | | | |
| Access and Confidentiality | | | | |
| 1) Is there any indication of any of the authors being involved with the running of the database or registry? | Yes | Yes | No | No |
| 2) Is there any indication that a Research Ethics Committee or Institutional Review Board or similar body has approved the research study? | No | No | No | (Yes) |
| 3) Is there any indication that patients have given informed consent to the research study? | No | No | No | No |
| 4) Are individuals' data records identifiable through a code or record ID number (regardless of whether the researchers can decode the information or not)? | Most likely, since this is single-centre | Most likely, since this is single-centre | Not necessarily | Yes |
| 5) Are any safeguards on confidentiality or data protection reported? | No | No | No | No |
| Routine data collection and processing procedures | | | | |
| 6) Is any routine data validation reported? | No | No | No | Yes |
| 7) Has any verification against alternative sources of data been reported on any data in the database or registry (either for the current or previous studies)? | No | No | No | Yes |
| 8) Does or did routine processing include manual copying / entry of data? | Most likely | Unsure | Unsure | Unsure |
| 9) Is the aim of the database or registry congruous with the aim of the study? | Yes | Yes | Unsure | No |
| 10) Number of centres involved in database / registry: | 1 | 1 | Unknown | 550 |
| 11) Are centre variations in clinical or reporting practices likely? | N/a | N/a | Yes | Yes |
| Assessment of data quality including validation | | | | |
| 12) Does the paper report that the included data have been assessed for consistency, accuracy, and completeness (either by database operators or researchers)? | No | No | No | No |
| 13) Do the researchers report having undertaken any assessment of data quality? | No | No | No | No |
| Individual variables | | | | |
| 14) Are the inclusion and exclusion criteria directly measured in the database or registry or any linked data source? | Yes | Yes | Yes | Yes |
| 15) Are the main outcome and exposure / intervention of interest directly measured in the database or registry or any linked data source? | Yes | Yes | Yes | Yes |
| 16) Are the identified predictors and confounders directly measured in the database or registry or any linked data source? | Yes | Yes | Yes | Yes |
| 17) If "no" to any of the previous three questions, have all used algorithms been validated? | N/a | N/a | N/a | N/a |

| | McDOUGALL | SEBALDT | TIEFEN-BRUNN | VAN STAA |
|---|---|---|---|---|
| **IV. ANALYSIS** | | | | |
| 1) Are estimates of the random variability of the main outcomes clearly described for each group of patients to be compared? | No | Yes | Yes | No |
| 2) Did the authors conduct appropriate statistical testing by: | | | | |
| i) Conducting statistical testing (when appropriate)? | Yes | Yes | Yes | Yes |
| ii) Reporting which statistical tests were used? | Yes | Yes | Yes | Yes |
| iii) Controlling for design effects in the statistical model? | Yes | Yes | N/a | Matching |
| iv) Controlling for repeated measures in populations that were followed over time? | No | No | N/a | No |
| v) Controlling for differential exposure to the intervention (or checking that it makes no difference)? | No | Yes | No | Partly |
| vi) Using a model designed to handle multi-level data when they included group-level and individual covariates in the model? | N/a | N/a | N/a | N/a |
| vii) Allowing for important centre differences in the analysis? | N/a | N/a | No | No |
| Missing or incomplete data | | | | |
| 3) Have patients with incomplete data been excluded? | Yes | No exclusions reported | No exclusions reported | No exclusions reported |
| 4) Have statistical adjustments been used for missing data? | No | Unsure | No | No |
| Multi-level data | | | | |
| 5) Have any measures of the overall adequacy of the statistical model been reported? | N/a | N/a | Yes | No |
| 6) Have any sensitivity analyses been performed? | Yes | No | No | Yes |
| **V. REPORTING** | | | | |
| 1) Was the study population well described? | Yes | No | Yes | Yes |
| 2) Was the intervention well described (what, how, who, where, how long/much)? | No | Yes | No | No |
| 3) Have the numbers of patients lost to follow-up been described? | Yes | Yes | N/a | Yes |
| 4) Was a power calculation reported? | No | No | No | No |
| 5) Are the main findings of the study clearly described? | Yes | Yes | Yes | CIs missing |
| 6) Have any explicit considerations of alternative explanations of the findings been presented? | No | No | Yes | Yes |
| 7) Are conclusions going beyond capabilities of the database or study design? | No | Unsure | No | No |

# DATA EXTRACTION FORM

Study ID

VERSION 13 Jan. 03
including alterations following pilot

Reviewer:

Date:

**PART 1: GENERAL INFORMATION**

**First author:** **Year of publication or report:**

**Title of publication or report:**

**Source of information:** Journal article ☐
Published or conference abstract ☐
Unpublished study ☐

**Detailed source (Journal, Conference, including detailed reference to source):**

**Affiliation (of first author) and contact address:**

**Countries participating:**

**Number of centres contributing data:**

**Stated main aim of the study:**

PART 2: PATIENT CHARACTERISTICS and SETTING

POPULATION CHARACTERISTICS

2.1 Source population and study setting (for each group):

| |
|---|
| |

2.2 Calendar years during which data included in study was collected:

| |
|---|
| |

STUDY SAMPLE

2.2 Specific study inclusion criteria:

Gender:

Age:

Others:

*2.3 Specific study exclusion criteria:*

Gender: Age:

Others:

2.4 Specific treatment and control conditions, including planned daily dose (mg):

| | |
|---|---|
| Treatment group A | |
| Treatment group B | |
| Treatment group C | |
| Treatment group D | |

2.5 Specific characteristics of sample at baseline:

| | Treatment group A | Treatment group B | Treatment group C | TOTAL | Significant difference? * |
|---|---|---|---|---|---|
| Number recruited: | | | | | |
| Male/Female numbers: | | | | | |
| Mean (SD)/ median (range)** age in years: | | | | | |
| Other characteristics reported: | | | | | |

* if "yes", record p-value ** underline which is entered

2.6 Duration of follow-up reported

| | |
|---|---|
| Planned (RCT only) | |
| Mean (SD) | |
| Median (range) | |
| Minimum | |
| Maximum | |

PART 3: INTERVENTION / "EXPOSURE"

3.1 Actual daily dose of drug, if different from planned regimen (in mg):

| Treatment group A | Treatment group B | Treatment group |
|---|---|---|
| | | |

3.2 Duration of drug treatment (observational studies only):

| | Treatment group A | Treatment group B | Treatment group |
|---|---|---|---|
| Mean (±SE/SD) | | | |
| Median (range) | | | |
| Minimum | | | |
| Maximum | | | |

### 3.3 Continuity of treatment and compliance:

| | YES | NO | UNCLEAR | N/A |
|---|---|---|---|---|
| Are ALL patients in all groups treated continuously? | | | | |
| Has compliance with treatment been assessed? | | | | |

## PART 4: OUTCOME MEASURES

### Which outcome measures are available?

| Outcome measures: | HOW measured? (definition of measure, instruments, person taking it, validity and reliability information) | WHEN measured? | | Same in all groups? |
|---|---|---|---|---|
| | | At baseline? | Post-baseline measurements (in weeks): | |
| PRIMARY OUTCOME: | | | | |
| OTHER OUTCOMES: | | | | |

## PART 5: ANALYSIS

| | YES | NO | UNCLEAR | N/A |
|---|---|---|---|---|
| 5.1 Intention to treat analysis used? | | | | |

**5.2 Number of patients included in study and excluded from analysis in each group:**

| | Treatment group A | Treatment group B | Treatment group | TOTAL | |
|---|---|---|---|---|---|
| | | | | No. | % |
| *Included at beginning of study:* | | | | | 100% |
| *Drop-outs before follow-up measurements* | | | | | |
| *Number of patients at final assessment* | | | | | |

**5.3 How is attrition dealt with in the analysis?**

**5.4 Reported reasons for dropouts**

***List in order of priority:***

| Treatment group A | Treatment group B | Treatment group |
|---|---|---|
| | | |

**5.5 Main analysis methods**

### 5.6 Variables controlled for in analysis:

Primary outcome analysis:

| Variables controlled for: | How was variable identified for control?* | Method of control:** |
|---|---|---|
| | | |

* e.g. bi-variate analysis, literature review, etc.
** e.g. restriction, randomisation, stratification, matching, multivariate analysis

Other outcome analysis:

| Variables controlled for: | How was variable identified for control?* | Method of control:** |
|---|---|---|
| | | |

* e.g. bi-variate analysis, literature review, etc.
** e.g. restriction, randomisation, stratification, matching, multivariate analysis

## PART 6: RESULTS

THIS ENTIRE SECTION HAD TO BE ALTERED SIGNIFICANTLY TO ACCOMMODATE SIMPLY THE OUTCOMES REPORTED IN A SPECIFIC OBSERVATIONAL STUDY - DIFFERENT FOR EACH CASE STUDY!

### 6.1 Mean (SD)* dose of corticosteroid use during study (mg):

| Treatment group A | Treatment group B | Treatment group |
|---|---|---|
| | | |

*indicate if other measure is reported (e.g. median)

6.2 Lumbar spine bone mineral density

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group A | Group B | p-value: | Difference between groups: | | |
| | | | | | | p-value: |
| Mean(SD/SE)* percentage change from baseline | | | | Mean ± SD | | |
| | | | | *Other:* | | |
| Mean (SD/SE)* absolute change from baseline ($g/cm^2$) | | | | | | |

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group | Group | p-value: | Difference between groups: | | |
| | | | | | | p-value: |
| Mean(SD/SE)* percentage change from baseline | | | | Mean ± SD | | |
| | | | | *Other:* | | |
| Mean (SD/SE)* absolute change from baseline ($g/cm^2$) | | | | | | |

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group | Group | p-value: | Difference between groups: | | |
| | | | | | | p-value: |
| Mean(SD/SE)* percentage change from baseline | | | | Mean ± SD | | |
| | | | | *Other:* | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mean (SD/SE)* absolute change from baseline ($g/cm^2$) | | | | | | |

*underline which is reported

6.3 Femoral neck bone mineral density

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group A | Group B | p-value: | Difference between groups: | | |
| | | | | | | p-value: |
| Mean(SD/SE)* percentage change from baseline | | | | Mean ± SD | | |
| | | | | *Other:* | | |
| Mean (SD/SE)* absolute change from baseline ($g/cm^2$) | | | | | | |

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group | Group | p-value: | Difference between groups: | | |
| | | | | | | p-value: |
| Mean(SD/SE)* percentage change from baseline | | | | Mean ± SD | | |
| | | | | *Other:* | | |
| Mean (SD/SE)* absolute change from baseline ($g/cm^2$) | | | | | | |

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group | Group | p-value: | Difference between groups: | | |
| | | | | | | p-value: |
| Mean(SD/SE)* | | | | Mean ± SD | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| percentage change from baseline | | | | *Other:* | | |
| Mean (SD/SE)* absolute change from baseline ($g/cm^2$) | | | | | | |

*underline which is reported

6.4 Trochanter bone mineral density

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group A | Group B | p-value: | Difference between groups: | | |
| | | | | | | p-value: |
| Mean(SD/SE)* percentage change from baseline | | | | Mean ± SD | | |
| | | | | *Other:* | | |
| Mean (SD/SE)* absolute change from baseline ($g/cm^2$) | | | | | | |

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group | Group | p-value: | Difference between groups: | | |
| | | | | | | p-value: |
| Mean(SD/SE)* percentage change from baseline | | | | Mean ± SD | | |
| | | | | *Other:* | | |
| Mean (SD/SE)* absolute change from baseline ($g/cm^2$) | | | | | | |

| YEARS FROM BASELINE: | | | | |
|---|---|---|---|---|
| | Group | Group | p-value: | Difference between groups: |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | p-value: |
| Mean(SD/SE)* percentage change from baseline | | | | Mean ± SD | | |
| | | | | *Other:* | | |
| Mean (SD/SE)* absolute change from baseline ($g/cm^2$) | | | | | | |

*underline which is reported

6.6 Fractures

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group A | Group B | p-value: | Difference between groups: | | |
| Number of patients (%) experiencing fractures | | | | | | p-value: |
| | | | | Relative risk±CI | | |
| | | | | Odds ratio±CI | | |
| Mean (SD/SE)* number of fractures per patient | | | | Mean ± SD | | |
| | | | | *Other:* | | |

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group A | Group B | p-value: | Difference between groups: | | |
| Number of patients (%) experiencing fractures | | | | | | p-value: |
| | | | | Relative risk±CI | | |
| | | | | Odds ratio±CI | | |
| Mean (SD/SE)* number of fractures per patient | | | | Mean ± SD | | |
| | | | | *Other:* | | |

| YEARS FROM BASELINE: | | | | | | |
|---|---|---|---|---|---|---|
| | Group A | Group B | p-value: | Difference between groups: | | |
| Number of patients (%) | | | | | | p-value: |
| | | | | Relative risk±CI | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| experiencing fractures | | | | Odds ratio±CI | | |
| Mean (SD/SE)* number of fractures per patient | | | | Mean ± SD | | |
| | | | | *Other:* | | |

*underline whichever is reported

NOTES

| |
|---|
| |